The Role of Micronutrients in Genetic Adaptation

Jasmin Rees
University College London
Institute of Child Health

Submitted to University College London (UCL) in partial fulfilment of the requirements for the awards of the degree of Doctor of Philosophy

Primary Supervisor: Professor Sergi Castellano
Secondary Supervisor: Dr Aida Andrés

Word Count: 71,717
April 2023

Dedicated to Sam and Sarah: for always telling me how smart I am, and for reminding me that I can also be incredibly stupid.

"Let's go girls" – Shania Twain

Declaration

I, Jasmin Rees, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis explores the role of micronutrients in genetic adaptation, largely focusing on micronutrients as the selective driver of local adaptation in modern humans. Additionally, the role of the micronutrient selenium in wider mammalian evolution is also investigated. Micronutrients are key dietary components in all organisms, needed in small, specific quantities and involved in a wide variety of essential metabolic processes. In modern humans, all micronutrients (with the exception of vitamin D) must be absorbed from the diet, since they cannot be synthesised within the body. Levels of dietary micronutrients in turn depend on the composition of the soil where plant and animal foodstuffs grow and feed, and hence can vary widely over different localities. As informed by a novel simulation framework, I use the allele-differentiation statistic F_{ST} and recently developed genealogical method *Relate* to identify signatures of natural selection in 40 diverse modern human populations in 276 genes associated with 13 micronutrients. I show signatures of positive selection are inferred in many global populations and micronutrient categories, and show that the strongest signatures of positive selection agree with known micronutrient composition of local soils and endemic deficiencies in modern human populations. I found no evidence for classic polygenic models of positive selection and infer that adaptation in response to micronutrients in the diet is most likely monogenic or oligogenic in nature. I evaluate the evidence for positive selection in genes associated with zinc, calcium, selenium, iron and iodine in detail and use a combination of methods to propose the origin and timing of selection acting on these micronutrient-associated genes. I propose that micronutrients are an important selective force in modern humans, and have shaped the genomic variation of our species. I also present the first evidence for molecular convergent evolution in mammalian proteins losing the selenium-containing amino acid selenocysteine for the sulphur-containing cysteine.

Impact Statement

There has been significant recent progress in understanding local adaptation amongst modern human populations, largely pertaining to methodological developments and increases in available genomic data, of both modern and ancient humans. Still, important questions and goals remain. This includes 1) evaluating the role of local adaptation, and respective selective drivers, in human genetic diversity and population differentiation; 2) evaluating the role of selection on standing variation in modern humans, which requires identification of subtle signatures of positive selection that can remain hidden in the genome; 3) addressing the current bias, at the time of writing, of studied populations, and including under-represented populations in studies of genetic diversity; and 4) identifying cases of local adaptation that have resulted in average phenotypic differences between populations in health-related traits.

In this thesis, I explore the role of dietary micronutrients in human local adaptation in the most comprehensive study to-date, identifying signatures of positive selection in 276 genes associated with 13 different micronutrients in 40 diverse populations. I propose dietary micronutrients as a key selective driver amongst modern human populations, building on previous literature and suggesting novel cases of micronutrient-associated adaptation in individual populations or regions. Here, I address point number 1) and point number 3). I also first identify two methods with increased power of identifying the subtle signatures of selection on standing variation, and identify potential instances of micronutrient-associated adaptation driven by selection on standing variation. This addresses point number 2).

Micronutrients play an essential role in human health, and understanding the interaction between micronutrient levels in the diet and genetic variation of diverse modern human populations is vital in understanding the differential risk of micronutrient deficiency amongst different populations. The signatures of positive selection identified here implies that different human populations may have, on average, different metabolic responses to varied dietary micronutrient levels and may therefore have increased risk of micronutrient deficiency or toxicity. Hence, I also address point number 4). The work presented here should thus prompt further study into the phenotypic consequences of such proposed adaptation, particularly under a changing dietary environment, made likely by changing soil levels under climate change and over-farming.

In this thesis, I also explore selenoprotein evolution in mammals, and propose a novel example of convergent adaptation leading to the development of novel function when a protein exchanges its catalytic residue selenocysteine to cysteine. This suggests the evolutionary pathway following the loss of selenocysteine can be narrow, and that other novel selenoprotein functions may remain currently hidden in nature.

Research Paper Declaration

Two publications are included in the writing of this thesis:

Rees, J., Castellano, S., Andrés, A., The Genomics of Human Local Adaptation, *Trends in Genetics,* 36, 6 (2020)

Gives the broad structure of ***Sections 1.4-1.6*** *(****Chapter 1****; Introduction).*

Rees, J., Sarangi, G., Cheng, Q., Floor, M., Andrés, A., Oliva Miguel, B., Villà-Freiza, Arnér, E., Castellano, S., Ancient loss of catalytic selenocysteine spurred convergent adaptation in a mammalian oxidoreductase, *Biorxiv*, doi: 10.1101/2023.01.03.522577, (2023)

Rewritten in the style of this thesis as ***Chapter 5****.*

Research Paper Declaration Forms given below.

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

1. **For a research manuscript that has already been published** (if not yet published, please skip to section 2)

   a) **What is the title of the manuscript?**

   The Genomics of Human Local Adaptation

   b) **Please include a link to or doi for the work**

   https://doi.org/10.1016/j.tig.2020.03.006

   c) **Where was the work published?**

   Trends in Genetics

   d) **Who published the work?** (e.g. OUP)

   Cell Press

   e) **When was the work published?**

   June 2020

   f) **List the manuscript's authors in the order they appear on the publication**

   Jasmin Rees, Sergi Castellano, Aida Andrés

   g) **Was the work peer reviewed?**

   Yes

   h) **Have you retained the copyright?**

   Yes (all authors)

   i) **Was an earlier form of the manuscript uploaded to a preprint server?** (e.g. medRxiv). If 'Yes', please give a link or doi)

No
If 'No', please seek permission from the relevant publisher and check the box next to the below statement:

☒

*I acknowledge permission of the publisher named under **1d** to include in this thesis portions of the publication named as included in **1c**.*

2. **For a research manuscript prepared for publication but that has not yet been published** (if already published, please skip to section 3)

3. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

Jasmin Rees: writing; Sergi Castellano: writing and editing; Aida Andrés: writing and editing

4. **In which chapter(s) of your thesis can this material be found?**

This work is used to broadly structure Sections 1.4-1.6 in Chapter 1 (Introduction).

5. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate
Click or tap here to enter text.
Date:
15/04/2023

Supervisor/ Senior Author (where appropriate)
Sergi Castellano
Date
16/04/2023

UCL Research Paper Declaration Form

referencing the doctoral candidate's own published work(s)

1. For a research manuscript that has already been published (if not yet published, please skip to section 2)

.

2. For a research manuscript prepared for publication but that has not yet been published (if already published, please skip to section 3)

a) **What is the current title of the manuscript?**

Ancient loss of catalytic selenocysteine spurred convergent adaptation in a mammalian oxidoreductase

b) **Has the manuscript been uploaded to a preprint server?** (e.g. medRxiv; if 'Yes', please give a link or doi)

https://doi.org/10.1101/2023.01.03.522577

c) **Where is the work intended to be published?** (e.g. journal names)

PNAS

d) **List the manuscript's authors in the intended authorship order**

Jasmin Rees, Gaurab Sarangi, Qing Cheng, Martin Floor, Aida Andrés, Baldomero Oliva Miguel, Jordi Villà-Freixa, Elias Arnér, Sergi Castellano

e) **Stage of publication** (e.g. in submission)

In submission

1. **For multi-authored work, please give a statement of contribution covering all authors** (if single-author, please skip to section 4)

   Jasmin Rees and Sergi Castellano wrote and edited; Jasmin Rees, Gaurab Sarangi, Aida Andrés and Sergi Castellano performed evolutionary analysis; Qing Cheng and Elias Arnér performed experiments; Baldomero Oliva Miguel, Jordi Villà-Freixa and Martin Floor performed theoretical modelling.

2. **In which chapter(s) of your thesis can this material be found?**

   Chapter 5 (Ancient loss of catalytic selenocysteine spurred convergent adaptation in a mammalian oxidoreductase).

3. **e-Signatures confirming that the information above is accurate** (this form should be co-signed by the supervisor/ senior author unless this is not appropriate, e.g. if the paper was a single-author work)

Candidate
Click or tap here to enter text.
Date:
15/04/2023

Supervisor/ Senior Author (where appropriate)
Sergi Castellano
Date
16/04/2023

Acknowledgements

It is very hard to start writing an acknowledgement section, and even harder to stop. I have had an amazing three and a half years working on this research, and I truly hope that some of that joy is conveyed on the pages here.

My first thanks are to my incredible supervisors, who will probably never quite realise how grateful I am for this experience and for all that they have taught me. Sergi, thank you for all your wise words (especially when you don't know you're saying them) and teaching me to be just a bit less British. Aida, thank you for sharing all the joy you find in your work and making sense of some of my worse explained ideas. If I attempt to list all my gratitude and all that I've learned from you both this **will** be the same length of the thesis, and quite frankly it's already long enough. I know I'll never stop looking back to all I've learned here.

Thank you to both research groups that I was fortunate to be a part of and learn from, and all the encouragement over the years. I am excited to move forward, but sad to leave behind people who have given me so much.

Thank you to my beautiful family, Mike, Elly, Julia, Sye and Pixel, for giving me such a loving (and beautiful!) environment to write up. I hope some of the Australian sunshine translates onto the page. Thank you to my parents and my brother, who have never doubted that I could do this.

To the entirety of Room 112, thank you for the gossip, hot girl walks and the endless free biscuits, and for making research fun even when it wasn't. Thank you to Aidan for all the free alcohol at conferences, and thank you to Carl for never telling me the embarrassing things I say after drinking it. Thank you to ICH for giving me a sister as well as this research opportunity. Carina, I refuse to ever learn what your research is about but it's probably excellent. So grateful that I know such incredible scientists and even better people.

I am so very lucky that my so many other friends (beautiful, talented and creative souls!) have listened to me waffle on about my little science world for so long now. These PhD years have been truly wonderful, and they wouldn't have been so without you all. Laura, Anna, Hettie, Yelena, Cesca, Eloise and Muneebah - I am so grateful to have such wonderful women with whom to celebrate life. You are all my inspiration and mean more to me than I could ever say. And to Rico, Rufus, Lewis, Jonas, Yvan, Hasan and Mark, you've really been true feminists and supported this woman in STEM.

My greatest thanks and lifetime of gratitude to my Birmingham friends. You have become the greatest family I could ever hope for and this thesis wouldn't exist without you, neither would the version of myself today. Sam, Tinaye, Keiron, Maybee, Will, Marco, Anna, Matt, Ella, Alex and, most of all, my dearest Sarah; I love you all so much, and I never feel luckier than when belting out auld lyne syne in the OKR basement. It may sometimes take a village, but sometimes it takes the best city in the world.

Table of Contents

Abbreviations

ABC Approximate Bayesian Computation

aDNA Ancient DNA

bp Base pairs

Cys Cysteine

DAF Derived Allele Frequency

dN Number of non-synonymous sites

dS Number of synonymous sites

GWAS Genome Wide Association Study

Kya Thousand years ago

MA-gene (sets) Micronutrient-associated gene (sets)

pMA-gene (sets) proxy-Micronutrient-associated gene (sets)

Mya Million years ago

SDN Selection on *de novo* mutation

Sec Selenocysteine

SNP Single Nucleotide Polymorphism

SSV Selection on Standing Variation

ZSCII-associated genes Zinc, Selenium, Calcium, Iron and Iodine-Associated Genes

Chapter 1: Introduction

1.1. Overview

Micronutrients are needed by all living organisms to maintain optimal fitness, whether that is by contributing to healthy growth and development, maintaining immunity or supporting key metabolic processes (Bhutta and Salam 2012; Bailey et al. 2015; Monteiro et al. 2015). Hence, micronutrients can drive genomic adaptations to regulate their metabolism, uptake or synthesis within the body (Herráez et al. 2009; Mariotti et al. 2012; White et al. 2015; Engelken et al. 2016; Roca-Umbert et al. 2022). Many micronutrients, including all but one of the micronutrients essential for human health, cannot be synthesised by the organism, and instead must be absorbed from the diet or directly from the soil (Hurst et al. 2013; Dhaliwal et al. 2019). The local environment (which shapes the local diet) may then directly affect or result in micronutrient-associated adaptation, which may differ on a large scale over different taxa, or even between populations of the same species (hereby referred to as local adaptation).

In this thesis I explore the role of micronutrients in genetic adaptation, with a particular focus on exploring local adaptation in modern humans in response to micronutrients levels in the diet, as well as the greater role of selenium in selenoprotein evolution. This chapter begins with a brief discussion on the various types of natural selection and the role they play in genome evolution (**Section 1.2**). The history of migrations and admixture of modern humans is then summarised to present appropriate context in which to consider how the signatures of local adaptation may present in different human populations (**Section 1.3**). The majority of this chapter then reviews our current understanding of local adaptation in modern humans, including common selective drivers (**Section 1.4**), the genomics of local adaptation (**Section 1.5**), and the current methods used to infer local adaptation events (**Section 1.6**).

Micronutrients as a specific driver of human local adaptation is then explored (**Section 1.7**), with discussion of endemic diseases associated with micronutrient deficiency and toxicity; the variation of micronutrient levels in different soils; and previously identified instances of adaptation in modern humans in response to micronutrient levels. Finally, the role of micronutrients in wider biology is briefly explored, before describing the evolution of selenoproteins in the context of a specific selenium-containing amino acid (selenocysteine; **Section 1.8**).

In **Chapter 2**, a simulation framework that models local adaptation in major human populations is used to test the power of different methods in identifying signatures of recent positive selection. The methods identified as having the highest power by these simulations (F_{ST} and $Relate$; (Weir and Cockerham 1984; Speidel et al. 2019) are then used in **Chapter 3** and **Chapter 4**.

In **Chapter 3**, the signatures of natural selection in genes (n=276) associated with 13 micronutrients within 40 diverse modern human populations are investigated. In **Chapter 4**, the adaptive signatures of genes associated with five micronutrients (zinc, calcium, selenium, iron and iodine) are further explored to suggest the potential origin and time of putative selection events, alongside the most likely selective drivers. Finally, in **Chapter 5**, the role of the micronutrient selenium in wider evolution is explored, by reconstructing the history and adaptive signatures surrounding the selenoprotein *GPX6* in mammals, which relies on selenium for catalysis. **Chapter 6** presents a summary of

this work, contextualising the findings across the field of human local adaptation and selenium biology.

1.2. Natural Selection

Natural selection is one of the four fundamental forces of evolution, alongside mutation, genetic drift and gene flow, and drives the evolution and persistence of adaptive phenotypes within a population. It was first formally defined by Charles Darwin in the mid-19th century (Darwin 1859), although also independently proposed by and developed alongside Alfred Russell Wallace (Darwin and Wallace 1858). It continued to be advanced by many biologists in the following century, most notably by Ronald Fisher and John Burdon Sanderson Haldane (Fisher 1919; Haldane 1924). Natural selection refers to the differential reproductive success of individuals within a population, whereby fitness advantage, or disadvantage, is conferred by underlying genotypes. In reference to natural selection, the fitness of a genotype is hence the relative fitness (ω; the absolute fitness relative to the fittest genotype). This serves as a measure of the relative fitness advantage or disadvantage of a genotype, which, within population genetics studies, is often referred to as the selection coefficient s.

The efficacy of natural selection depends on this relative fitness (s) but also on the magnitude of genetic drift (Hahn 2018). In populations with a small effective population size (N_e, the value that represents the size of an idealised Wright-Fisher population showing the same loss of genetic diversity, and is usually lower than the census population size (Hahn 2018), the effect of genetic drift is larger, and can cause a random subset of genetic variants to rise to high frequencies, irrespective of their selection coefficient. The action of natural selection is also heavily interlinked with the remaining evolutionary forces, gene flow and mutation, both of which introduce new genetic variants into a population (where mutations are the only source of true novel genetic variants). Therefore, natural selection must be considered alongside such forces. In humans, the specie of interest in **Chapters 2-4**, the role of highly diverse demographic and migration histories across populations (see **Section 1.3**) in the efficacy of natural selection must be considered. More than this, complex histories of migrations and gene flow must be recognised when considering how genetic diversity may be a result of either natural selection or neutral processes (see **Section 1.5**).

Natural selection also underpins the Modern Synthesis evolutionary theory (Huxley 1942) which reconciled Darwin's concept of natural selection with a population-oriented view of Mendelian genetics. A key idea here is that whilst natural selection acts on all individuals, it results in evolution at the population level (where evolution is most explicitly the change in frequency of alleles within a population). Whilst this theory has since been further developed, including expanding our view of inheritance to not solely gene-based but also epigenetic or cultural inheritance (Laland et al. 2015), it still remains a key framework of which to understand evolution and adaptation.

1.2.1. Selection in Genome Evolution

Both natural selection and neutral evolutionary processes drive the fate of mutations following their appearance within a population. It is important to consider how natural selection ultimately contributes to overall genome evolution, and how the role of N_e and mutation fitness may contribute to this.

Broadly speaking, mutations can be categorised as advantageous, deleterious or neutral, where some weakly deleterious or advantageous mutations may be referred to as nearly-neutral should they exist in populations dominated by genetic drift (*i.e.,* low population size; with the upper limit to effective neutrality approximately $|N_e s| \approx \frac{1}{4}$ (Loewe and Hill 2010)). The majority of non-neutral mutations are strongly deleterious (Keightley and Eyre-Walker 2010; Trindade et al. 2010), and are rapidly purged from populations via purifying selection. Intuitively, it then remains that the majority of the observed genetic variation is largely a product of advantageous and neutral mutations (where the latter includes weakly deleterious and weakly advantageous mutations that behave as neutral when N_e is particularly low).

Prior to the 1960s, genome evolution was believed to be primarily driven by positive selection: selection that increases the frequency of advantageous alleles faster than what would be expected under neutral drift (Sabeti et al. 2006). Existing polymorphisms were then believed, although with some contention (Asthana *et al,* 2005), to reflect balancing selection: selection maintaining multiple alleles within a population or species, thereby driving advantageous genetic and phenotypic diversity (Andrés 2011; Bitarello et al. 2018). Implicitly, this represented the view that genetic differences between populations and species were mostly reflecting their own adaptive processes (Duret 2008).

However, with the availability of sequence data came the birth of the seminal neutral theory of evolution, most notably proposed by Mottoo Kimura (Kimura 1968). This theory states that the vast majority of evolutionary changes at the molecular level are not in fact a product of positive selection, but rather random fixations of selectively neutral or nearly neutral mutations. Hence, this makes two key points about how natural selection affects genome evolution: 1) the great majority of molecular differences between species is due to nearly neutral substitutions and 2) polymorphic alternative alleles within species have neutral fitness effects with respect to each other (with their dynamics dominated by mutation-drift equilibrium) (Hahn 2018).

Whilst this model remains arguably the most dominant explanation of molecular variation, at both the level of intraspecific genetic variation and interspecific genetic divergence, and is often used as a null model representing the most parsimonious scenario across species (Hahn 2018), other models have since been proposed to fully represent observed patterns of variation. This includes the nearly neutral theory of molecular evolution which emphasises the role of slightly deleterious mutations (Ohta 1973, 1976), as well as models that emphasise the role of selection on linked neutral variation, either purifying selection (background selection model; (Charlesworth et al. 1993, 1995)) or positive selection (hitchhiking model; (Smith and Haigh 1974; Kaplan et al. 1989)). Whilst these models do exhibit key differences, they are united in their assumption that the majority of polymorphisms are not maintained by positive selection.

1.2.2. Adaptive Convergence

Positive selection may not be as pervasive in genomes as first thought, but it remains the driving force behind the prevalence of advantageous traits across species. Whilst many adaptations are unique to species or populations (Sabeti et al. 2006; Savolainen et al. 2013), some adaptive phenotypes may be shared although independently acquired.

This is adaptive convergence: repeated acquisition of the same phenotype from independent lineages (Storz 2016; He et al. 2020).

Adaptive convergence may be caused by both convergent or parallel changes at the amino acid level (Storz 2016; He et al. 2020). That is to say, the same phenotypes may be acquired from substitutions at a particular site from different ancestral amino acids to the same derived amino acid (convergent substitution) or from sites that have independently changed from the same ancestral amino acid to the same derived amino acid (parallel substitution, more common in closely related species) (see **Fig. 1.1**). In an extension of this, substitutions at different sites, or indeed variants of different genes, may even confer the same adaptive phenotype (Witt and Huerta-Sánchez 2019).

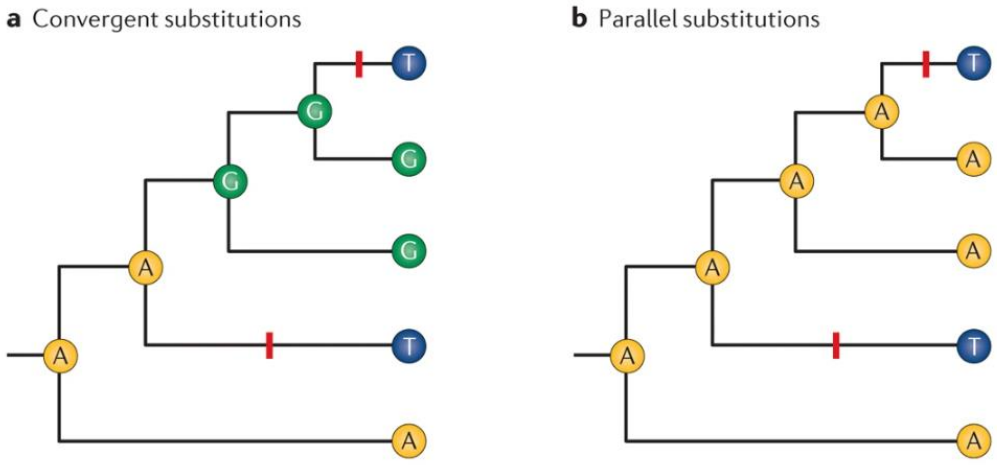


Figure 1.1: Convergent and parallel amino acid changes. *Convergent substitutions from a different ancestral amino acid to the same derived amino acid (**a**) and parallel substitutions from the same ancestral amino acid to the same derived amino acid (**b**) may lead to acquisition of the same adaptive phenotype. Taken from (Storz 2016).*

This variation in acquisition of the same adaptive phenotype reflects the many-to-one mapping of genotype to phenotype. However, not all genotypes mapping to the same phenotype necessarily have the same probability of fixation. Probability of fixation of advantageous mutations once they arise is not only dictated by the strength of selection and the demography of the population they are in, but by the role of pleiotropy, the phenomenon where a single mutation or genetic locus affects multiple traits (Solovieff et al. 2013). For a given set of possible mutations that result in the same phenotypic response, those which have the lowest degree of deleterious pleiotropy, *i.e.,* those that have the least deleterious effect on other genetically related functions, have higher fixation probability (Storz 2016).

Epistasis, or the functional interaction between genes (Phillips 2008), must also be considered in adaptive convergence. This phenomenon describes how different mutations have varying fitness depending on the underlying genetic background. Thereby epistasis also narrows the set of possible mutations that may respond similarly to selection (Lunzer et al. 2010; Storz 2016), an effect particularly strong in more divergent species. This is not necessarily independent from the effects of pleiotropy: mutations that can compensate for the reduction of fitness arising from deleterious pleiotropy also depend on the genetic background (Solovieff et al. 2013; Storz 2016).

1.3. Genetic History of Modern Humans

1.3.1. Major Migrations in Human History

Modern humans differ from many other species in that they inhabit almost all areas of the globe. The history of modern humans is therefore tightly interwoven with a series of large- and small-scale migration events, which can vary drastically over different populations. These migration events have profoundly affected the genomic variation across human populations and therefore must be considered in parallel with the genomic effects of selection (see **Section 1.5**).

Modern humans originated in Africa less than 200kya years ago (Reich et al. 2010), and are identified in the fossil record by a wealth of anatomical traits, particularly those defined in the crania: a high frontal bone, weak supraorbital torus and small dentition with canine fossa (Stringer and Andrews 1988; White et al. 2003). The development of sequencing technology resulted in genetic evidence for an African origin of modern humans, such as seminal studies showing mitochondrial and Y-chromosome haplotypes as subsets of those identified within Africa (Soares et al. 2012; Haber et al. 2019) as well as those demonstrating a decrease in genetic diversity as a function of geographic distance from East or South Africa (Prugnolle et al. 2005; Ramachandran et al. 2005). Still, debates remain on the exact region of Africa where modern humans evolved. Historically, an East African origin has often been suggested, supported by the oldest anatomically modern human fossils in Ethiopia (Clark et al. 2003; McDougall et al. 2005; McCarthy and Lucas 2014). However, genetic studies have suggested a variety of origins, including East Africa as well as Southern and Northern regions (Henn et al. 2011; Blome et al. 2012; Schlebusch et al. 2012; Fadhlaoui-Zid et al. 2013). Still, these studies are united in their acknowledgement that the complex population history of the continent complicates any such conclusion.

The Out of Africa migration (Stringer and Andrews 1988), or the event that resulted in modern humans colonising the world outside of the African continent, is therefore a substantial migration event in human evolutionary history. That is not to say migrations only occurred outside of Africa; Africa itself has a rich and varied migration history, although on the whole is less resolved than the migrations of non-African regions (due to increased complexity, lower availability of modern and ancient genomes, and historical bias). Some notable African migrations include the Bantu expansion (beginning ~5kya at the Cameroon and Nigeria border, eastward to Uganda at ~3kya and then southwards to Mozambique and South Africa at ~1.8kya and ~1.5kya, respectively (Beltrame et al. 2016)), the spread of pastoralism into sub-Saharan Africa (Afroasiatic populations migrating from Ethiopia into Kenya and Tanzania ~5kya (Patin et al. 2017)) and bidirectional migrations through the Sahel (between east and west Africa over the past ~8ky (Hirbo et al. 2012)).

Whilst the exact dynamics of the Out of Africa migration are still debated (as more comprehensively reviewed in (López et al. 2016)), the majority of genetic evidence places the date of the main body of this migration as approximately 60kya (Zhivotovsky et al. 2000; Underhill and Kivisild 2007; Shi et al. 2010; Gravel et al. 2011; Harris and Nielsen 2013). Some archaeological evidence, including modern human teeth identified in Southern China (~80-120kya; (Liu et al. 2015)) and Australian modern human fossils (~56-40kya; (Bowler et al. 2003)), suggests an earlier migration through to East Asia and Oceania (Armitage et al. 2011; Rose et al. 2011). It is important to note that not all

migrations of modern humans necessarily resulted in descendant extant populations. Indeed, the early modern human fossils found in the Levant (dated 120-90kya; (Grün et al. 2005)) are suggested to be a result of an earlier "failed" exodus from Africa (Pope and Terrell 2007; Hershkovitz et al. 2015; Kuhlwilm et al. 2016).

There are two main suggested routes that modern humans took on the exit from Africa: the Northern route (through Egypt and Sinai (Luis et al. 2004; Pagani et al. 2015)) and the Southern route (through Ethiopia, the Bab el Mandeb strait, and the Arabian Peninsula; (Quintana-Murci et al. 1999; Fernandes et al. 2012; Soares et al. 2012) see **Fig. 1.2**). However, there remains no confident consensus from either genetic or archaeological evidence on which route was taken (López et al. 2016). Some genetic evidence supports migrations at different timepoints possibly along both routes, with both migrations resulting in descendant populations today (Lahr and Foley 1994; Rasmussen et al. 2011; Reyes-Centeno et al. 2014; Tassi et al. 2015). In particular, this supports the hypothesis that Southeast Asian and Oceanic populations are the descendants of a first migratory wave Out of Africa estimated as approximately 70-100kya (Rasmussen et al. 2011; Reyes-Centeno et al. 2014, 2015; Tassi et al. 2015).

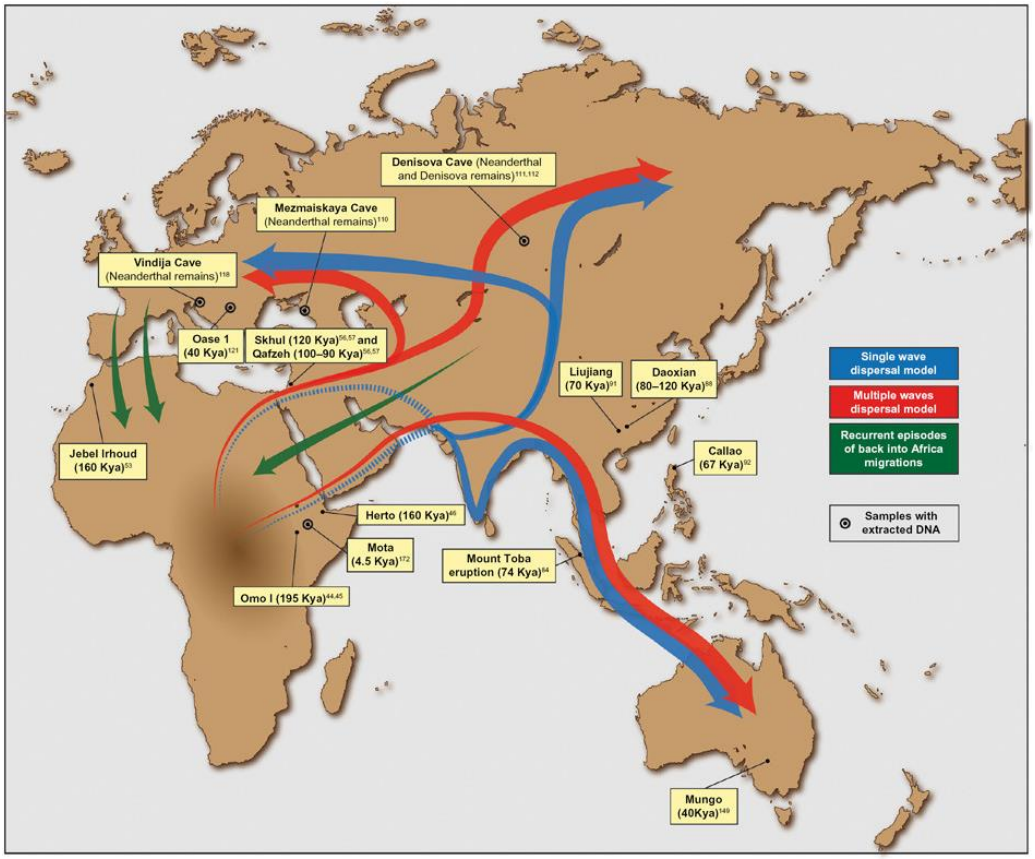


Figure 1.2: Overview of modern human migrations out of Africa. *Putative migration waves Out of Africa, and following migrations, are shown according to various models. Significant human remains and archaeological sites also given. Taken from (López et al. 2016).*

All such migrations described here, including the Out of Africa migration, remain estimations rather than known events, but do represent the general consensus in the field, although not without uncertainties (particularly surrounding the exact timing of migration or admixture events). Still, the Out of Africa migration refers to the strongly supported migration at around 60kya of a population ancestral to modern Eurasians and Americans (Zhivotovsky et al. 2000; Underhill and Kivisild 2007; Shi et al. 2010; Gravel et al. 2011; Harris and Nielsen 2013). This ancestral population then expanded across Eurasia, resulting in a spatially structured modern human population across this region by 40kya (Fu et al. 2014; Seguin-Orlando et al. 2014). East Asians are also suggested to have then received gene flow from populations ancestral to Aboriginal Australians already having colonised Asia following the proposed earlier migration (Rasmussen et al. 2011). Approximately 20kya, a population descended from East Asians with substantial north Eurasian gene flow migrated to the Beringian strait, before migrating downwards into the Americas, eventually giving rise to northern and southern Native American populations (Raghavan, Skoglund, et al. 2014; Rasmussen, Anzick, et al. 2014; Moreno-Mayar et al. 2018). Naturally, with such a migration, or multiple migrations as some work has suggested (including migrations harbouring Austronesian ancestry (Skoglund et al. 2015)), modern American population history is characterised by an extreme population bottleneck (Prugnolle et al. 2005; Ramachandran et al. 2005; Gravel et al, 2013; Fagundes et al, 2018).

The time following the Neolithic transition (10-5kya (Hawkes 1949)) resulted in multiple population movements and subsequent gene flow, particularly well documented across Europe. The most notable migrations include that of Anatolian farmers of the Near East into early western European populations (Haak et al. 2010; Skoglund et al. 2012; Lazaridis et al. 2014). The admixture between this group and the hunter-gatherer groups already present in Europe resulted in what is termed the Early European Farmer population (Mathieson et al. 2015a). This population was later largely replaced by a population associated with the Yamnaya culture, who migrated into western Europe from the steppe of Eastern Europe surrounding 3000BC and from which modern European ancestry is mostly derived (Allentoft et al. 2015; Mathieson and Terhorst 2022).

The migratory history of modern humans given here is not designed to be fully comprehensive, but to give an overview of the range of population histories of the human species. It also must be highlighted that population histories are not independent from each other following divergence events, as back-migrations were likely common throughout human history (González et al. 2007; Moreno-Mayar et al. 2018). In particular, African and non-African populations are not independent in the time following the Out of Africa migration(s); migrations from Eurasian populations back into some African populations are believed to have resulted in the high levels of non-African ancestry in some regions of Africa (Maca-Meyer et al. 2001; González et al. 2007; Pagani et al. 2012; Pickrell et al. 2014). The number and timing of these migrations remain a question, but must also be considered when exploring African genetic diversity (López et al. 2016).

1.3.2. Introgression with Archaic Humans

In recent years, it has become clear that there has been extensive admixture with archaic humans in the evolutionary history of anatomically modern humans (Green et al. 2010; Reich et al. 2010; Meyer et al. 2012; Prüfer et al. 2014). Archaeological

evidence has long placed Neanderthals across Eurasia (Higham et al. 2014), with more recent archaeological data placing their sister group Denisovans in the same localities of Siberia (Reich et al. 2010). With the development of ancient DNA sequencing technology and analysis has come substantial evidence for gene flow events between these archaic and modern human populations (see **Figure 1.3**; (Green et al. 2010; Reich et al. 2010; Meyer et al. 2012; Castellano et al. 2014; Prüfer et al. 2014; Reilly et al. 2022)).

These gene flow events, or introgression events, did not occur between all modern and archaic populations, and instead were more localised events that have resulted in different proportions of archaic DNA in the genomes of various modern human populations. Non-Africans have approximately 1.5-2.1% of DNA of Neanderthal origin, with this proportion slightly higher in Asian individuals compared to Europeans (Green et al. 2010; Prüfer et al. 2014). Historically, there was the suggestion that this higher proportion in East Asians was due to a following bottleneck and greater genetic drift in East Asian population history, rather than a separate pulse of introgression (Keinan et al. 2007; Skoglund and Jakobsson 2011; Sankararaman et al. 2012). However, more recent studies have implied a far more complex history of multiple introgression events across ancient European and Asian populations which may explain these differences (Villanea and Schraiber 2019; Iasi et al. 2021; Schaefer et al. 2021).

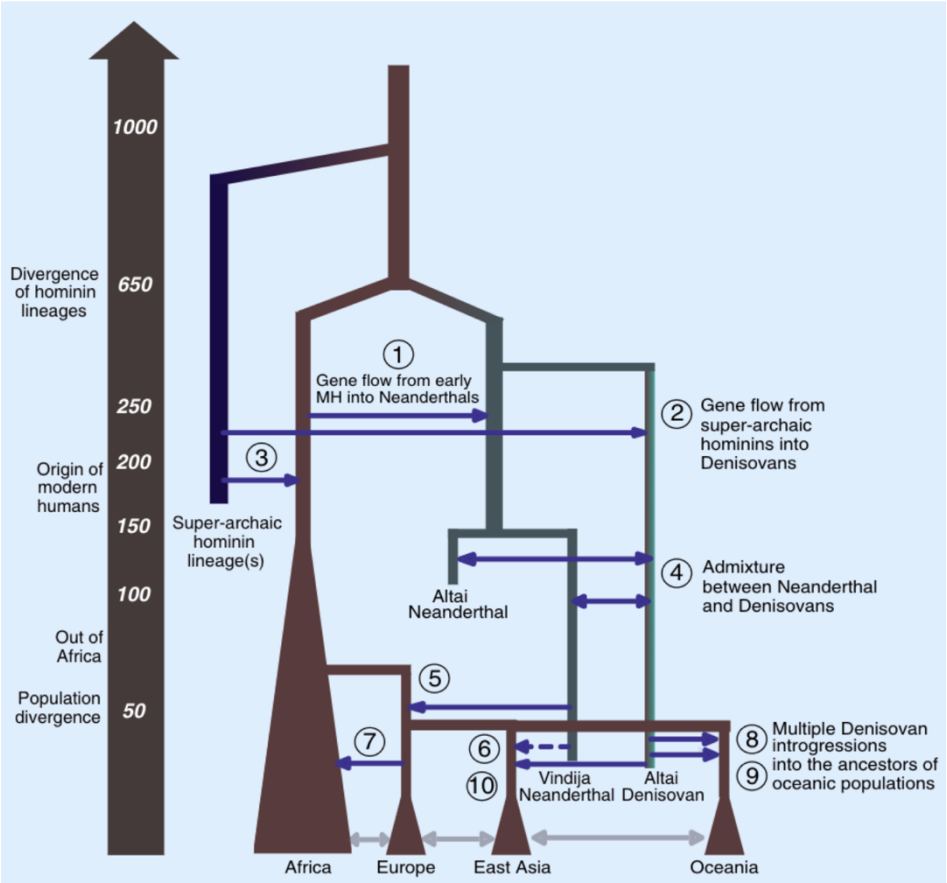


Figure 1.3: Overview of modern human demographic history. *A simplified overview of the demographic history and inferred gene flow events between modern and archaic humans, including un-named "ghost" populations. Taken from (Reilly et al. 2022).*

By measuring the length of inferred introgressed tracts of DNA, the timing of the introgression events from Neanderthals into Eurasians is estimated to be approximately 50-60kya, suggested to have occurred in the region of the Middle-East (Sankararaman et al. 2012; Fu et al. 2014; Seguin-Orlando et al. 2014). This estimated introgression time supports introgression rather than ancestral population structure as the cause of Neanderthal ancestry in only non-Africans, which has been proposed as a counter-explanation. However, if ancient population subdivision in a common ancestral population of archaic and modern humans drove the observed Neanderthal ancestry in Eurasians, it would be expected that these tracts would be dated more closely to the time of Neanderthal-modern human divergence (estimated older than 200kya (Prüfer et al. 2014)); this is hence discounted by the majority of the field and Neanderthal introgression remains well supported.

Denisovan ancestry, is predominantly found in Melanesians, Papuans and Australians at higher proportions of 3-6% (Reich et al. 2010, 2011; Meyer et al. 2012), alongside smaller proportions found in East Asians (0.2%) (Skoglund and Jakobsson 2011; Browning et al. 2018). Similar to what is now understood to be the case in Neanderthal-modern human introgression, it appears that introgression events between Denisovans and modern humans occurred multiple times in human evolutionary history (Browning et al. 2018; Jacobs et al. 2019; Schaefer et al. 2021). At least three separate Denisovan lineages have been inferred to contribute to modern human genetic variation, all of which appear divergent from each other and likely represent geographically separated archaic populations (Jacobs et al. 2019).

Finally, more recent studies have suggested that currently unknown archaic human groups have contributed to the genomes of contemporary populations, including African populations (Mondal et al. 2019; Wall et al. 2019; Durvasula and Sankararaman 2020; Hubisz and Siepel 2020; Wang et al. 2020). These ghost introgression events are between modern humans and populations for which we currently have no genomic data, and therefore remain in most senses unresolved. Still, this not only implies that modern humans lived contemporaneously and coexisted with multiple groups, but also suggests a considerably deeper and interwoven history of admixture between modern, archaic and potentially super-archaic humans (Ahlquist et al. 2021).

1.4. Local Adaptation in Modern Humans

Local adaptation is defined as when, due to genetic differences, individuals from a population have a higher average fitness in their local environment than those from other populations of the same species (Rees et al. 2020). This occurs when populations are exposed to different selective pressures, often tightly related to local environment (Savolainen et al. 2013; Tiffin and Ross-Ibarra 2014). Ultimately, this population-specific natural selection results in genetic and phenotypic differentiation between populations.

Local adaptation in modern humans is of particular interest because, as a species, we inhabit almost all environments across the globe, including some of which are considered extreme (Ilardo and Nielsen 2018). Moreover, following the Out of Africa migration (Soares et al. 2012; López et al. 2016; Haber et al. 2019), humans have colonised many of these environments rather rapidly (see **Section 1.3.1**), with novel environmental conditions expected to exert potentially strong selective pressures.

Environments within Africa are also incredibly diverse and have the potential to result in local adaptation events (Fan et al. 2023).

1.4.1. Common Selective Drivers

Many well documented examples of local adaptation within humans exist, and are in strong support of local adaptation contributing to modern human genetic variation and differentiation between populations (see **Table 1.1;** (Rees et al. 2020)). These adaptations are most commonly shown to be in response to diverse diets (Perry et al. 2007; Tishkoff et al. 2007a; Schlebusch et al. 2012; Fumagalli et al. 2015; White et al. 2015; Minster et al. 2016; Evershed et al. 2022), pathogens (Fumagalli et al. 2011; Karlsson et al. 2014; Nédélec et al. 2016; McManus et al. 2017), elevation (Yi et al. 2010; Bigham and Lee 2014; Huerta-Sánchez et al. 2014) and ambient temperature (Key et al. 2018), as well as to mediate local cultural pressures, such as breath-hold diving in the Bajau (Ilardo et al. 2018). Other local selective pressures were driven by the Neolithic revolution and development of agriculture, and include dietary changes as well as increased pathogen risk that accompanied densely packed living conditions and exposure to zoonotic diseases (Latham 2013; Domínguez-Andrés et al. 2021).

In some cases, these adaptations are convergent in nature, with different populations developing adaptive phenotypic responses to the same environmental pressure via different genes. This is most notable in the adaptation to high altitude in Ethiopian, Andean and Tibetan populations (Bigham and Lee 2014; Witt and Huerta-Sánchez 2019), light skin adaptation in Europeans and East Asians (Norton et al. 2007), and the adaptation allowing consumption of milk past weaning, independently conferred by multiple different variants upstream of the *LCT* gene in African and European populations (Tishkoff et al. 2007a; Evershed et al. 2022). However, the adaptive function of other convergent phenotypes remains unclear. For example, the small-stature phenotype (mean adult height below 152cm) is a characteristic trait of multiple rainforest hunter-gatherers, living in Central Africa, South America and Southeast Asia (Perry and Dominy 2009). This trait appears to be under selection, and driven by a currently unknown, and debated, selective pressure (Herráez et al. 2009; Perry and Dominy 2009; Venkataraman et al. 2018).

Table 1.1: Overview of the known genes under local adaptation in human populations and their proposed selective pressures*. [a]: indicates gene variants that are a result of adaptive introgression with Neanderthal populations [b]: indicates gene variants that are a result of adaptive introgression with Denisovan populations [c]: selection acting on* ***structural variations*** *(deletions, insertions, inversions, duplications' and copy number variations). Selection noted as across populations indicates that selection is seen differentially across multiple populations according to the strength of the selective pressure. Taken from (Rees et al. 2020).*

Category	Selective Pressure	Gene Target(s)	Population with Adaptations	Refs
Diet	Lactose Post-weaning	*LCT*	Eurasians and Africans	(Bersaglieri et al, 2004; Sabeti et al. 2007; Tishkoff et al. 2007)
	Fatty Diets	*FADS*	Greenland Inuit	(Fumagalli 2015)
	High Arsenic levels	*AS3MT*	Argentinians	(Schlebusch et al. 2015)
	Low Selenium levels	*DI2, SelS, GPX1, GPX3, CELF1, SPS2, SEPSECS*	Chinese	(White et al. 2015; Engelken et al. 2016; Davy and Castellano 2018)
	Low Iron levels	*HFE*	Europeans	(Toomajian et al. 2003; Heath et al. 2016a)
	Low Iodine levels	*TRIP4*	Central African Pygmies	(López Herráez et al. 2009)
	Low Calcium levels	*TRPV6*	Non-Africans	(Hughes et al. 2008; Kovacs et al. 2013a)
	Zinc Levels	*SLC30A9, SLC39A8*	East Asians and Africans	(Zhang et al. 2015a; Engelken et al. 2016)
	Ergothioneine deficiency	*IBD5 (SLC22A4, SLC22A5)*	Europeans	(Wang et al. 2012)
	Frequent Starvation	*CREBRF*	Samoans	(Minster et al. 2016)
	Alcohol Consumption	*ADH1B*	Asians	(Osier et al. 2002; Han et al. 2007; Li et al, 2007)
	Starchy foods	*AMY1*[c]	Across populations	(Perry and Dominy 2009)
Pathogens	Malaria	*HBB, HBA, HPA, GYPA, GYPB, GYPC, G6PD, FY*	Sub-Saharan Africans	(Kwiatkowski 2005; McManus et al. 2017; Pierron et al. 2018)
	"African Sleeping Sickness"	*APOL1*	Western Africans	(Genovese et al. 2010)
	Hepatitis C	*IFNL4, IL28B*	Eurasians	(Ge et al. 2009; Key et al. 2014; Lu et al. 2014)
	HIV	*CUL5, TRIM5, APOBEC3G*	Biaka	(Ge et al. 2009; Zhao et al. 2012; Lu et al. 2014)
	General pathogen load	*ADAM17, ITGAL, LAG3, IL6, LRRC19, PON2, OAS1*[b]*, OAS* group[a]*, HLA* group[a]*, STAT2*[b]*, STAT2*[a]*, TLR1-TLR6-TLR10*[a]	Across populations	(Abi-Rached et al. 2011; Fumagalli et al. 2011; Mendez et al. 2012a, 2012b, 2013; Fumagalli 2015; Mathieson et al. 2015a; Nédélec et al. 2016)
Oxidative Stress	High Altitude	*EGLN1*	Andeans, Tibetans	(Bigham et al. 2010; Simonson et al. 2010; Bigham and Lee 2014)
		EPAS1[b]	Tibetans, Han Chinese	(Yi et al. 2010; Bigham and Lee 2014; Huerta-Sánchez et al. 2014)
		VAV3, ARNT2, THRB	Ethiopians	(Scheinfeldt et al. 2012)
	Breath-Hold Diving	*PDE10A, BDKRB2*	Bajau (Indonesia)	(Ilardo et al. 2018)
Cold Resistance	Cold Temperature	*TRPM8*	Eurasians	(Key et al. 2018)
	Polar Diet	*CPT1A, LRP5, THADA*	Siberians	(Cardona et al. 2014)
		PRKG1	Siberians	(Cardona et al. 2014)
		TBX15	Greenlandic Inuit	(Fumagalli 2015)
UV Exposure	Low UV levels	*SLC24A5, SLC45A2, OCA1-4, TYRP1, DCT, TYR, MC1R*[a]*, HYAL2*[a]	Across populations	(Nakayama et al. 2006; Edwards et al. 2010; Hancock, Alkorta-Aranburu, et al. 2010; Paschou et al. 2010; Yang, Novembre, et al. 2012)
	Low Vitamin D levels	*DHCR7, NADSYN1*	Northern European populations	(Mathieson et al. 2015a)
Height	Undetermined	*DOCK3, CISH, HESX1, POU1F1*	Central African rainforest hunter-gatherers	(Perry and Dominy 2009; Jarvis et al. 2012; Lachance et al. 2012)
	Undetermined	Highly polygenic	Europeans	(Turchin, Chiang, Palmer, Sankararaman, Reich, Hirschhorn, et al. 2012;

				Berg and Coop 2014; Mathieson et al. 2015a; Field et al. 2016; Sanjak et al. 2018; Berg, Zhang, et al. 2019; Sohail et al. 2019)
Assorted[1]	Undetermined	*EDAR*	East Asians	(Sabeti et al. 2007; Grossman et al. 2010; Adhikari et al, 2015; Reyes-Reali et al, 2018, Kataoka et al, 2021)
Unknown	Undetermined	17q.21.31 gene region[c]	Icelandic	(Stefansson et al. 2005)

1.4.1.1. Dietary Adaptation

Diet is arguably one of the most notable local selective pressures in humans, with lactase persistence often identified as representing the strongest signature of selection in Eurasian populations (Mathieson et al. 2015a; Speidel et al. 2019; Evershed et al. 2022). However, the diversity of diets across the globe is represented by more than the milk-drinking habits of certain populations. Some other notable differences in diet are due to underlying environmental conditions or food availability, whilst others are more tightly associated with cultural or societal developments.

Environments affect the human diet in many ways: not only the surrounding availability of plant and animal foodstuffs, but also the local soil composition (which affects the nutrient content in consumed plant and animal matter (Alloway 2013)). Indeed, adaptation in response to diet caused by environmental factors has been inferred in modern humans. This includes adaptation to frequent periods of starvation in Samoan populations (Minster et al. 2016), adaptation to the high fatty acid content found in Arctic diets (Fumagalli et al. 2015) and various proposed adaptations to deficient or toxic levels of trace minerals in local soils (Herráez et al. 2009; Schlebusch et al. 2012; White et al. 2015; Zhang et al. 2015a). These adaptations in response to trace mineral levels include those of micronutrients essential to the human diet (Shenkin 2006; Tulchinsky 2010) (see **Section 1.7**), such as selenium, iodine and zinc (Herráez et al. 2009; White et al. 2015; Zhang et al. 2015a), as well an adaptive response to toxic levels of arsenic (Schlebusch et al. 2015).

Dietary culture has also been shown to result in local adaptations, particularly following the agricultural revolution (which resulted in many changes to modern human life including vastly different diets (Diamond 2002; Naugler 2008; Latham 2013; Domínguez-Andrés et al. 2021; Evershed et al. 2022)). The increased copy number of the *AMY1* gene in various populations has been proposed as an adaptive response to increased amounts of starchy foods in agricultural diets (Perry et al. 2007) (although there has been some debate on the accuracy of determining the copy number of *AMY1* in this study (Ooi et al. 2017)) and lactase persistence is an adaptation associated with the practice of drinking milk post weaning (Tishkoff et al. 2007a; Evershed et al. 2022). Deficiencies of alcohol and aldehyde dehydrogenase, which result in reduced alcohol metabolism and risk of alcoholism, has also been suggested to be adaptive in some way (Osier et al. 2002; Han et al. 2007, Li et al. 2007), potentially associated with fermentation practices that arose following agriculture.

[1] A derived *EDAR* variant is associated with thicker hair, tooth and ear shape, sweat gland density and chin protrusion (Fujimoto et al, 2008; Adhikari et al, 2015; Reyes-Reali *et al*, 2018, Kataoka et al, 2021).

1.4.2. Local Adaptation and Health

The history of recent divergences and subsequent admixture amongst human populations means that the majority of genetic and phenotypic variation is found within, rather than between, populations (Rees et al. 2020). Still, local adaptation has resulted in average phenotypic differences amongst populations. Many of these adaptations are directly relevant to the health of contemporary populations, and result in population differences in the genetic risk or prevalence of disease.

In some cases, this is due to evolutionary mismatch, or a previously advantageous and selected trait becoming deleterious in contemporary environments (Manus 2018). The most notable example of this is seen in Samoan populations, who have a high frequency of the *CREBF* gene variant that allows rapid weight gain. This variant was likely beneficial under frequent starvation conditions, but now, in a modern state of food abundance, increases the risk of type 2 diabetes and related metabolic disorders (Minster et al. 2016). Similar issues are seen in Canadian, Greenlandic Inuit, and Siberian populations with a particular *CPT1A* gene variant. This variant maintains sugar homeostasis during a similarly nutrient sparse environment (low carbohydrate intake), but is now associated to hypoketotic hypoglycaemia and high infant mortality (Clemente et al. 2014).

Migrations and ecological change may also expose modern populations to novel environmental conditions to which they have not previously adapted. Relevant environments here are those that drove the suggested adaptations to trace mineral levels in the soil (*e.g.,* toxic arsenic levels and low selenium levels (Schlebusch et al. 2015; White et al. 2015)). Individuals that migrate to these environments and lack these genetic adaptations may face numerous health issues if the respective toxicity or deficiency is not addressed via other means. Similarly, individuals with such genetic adaptations who migrate from these environments to other geographic regions may also be more susceptible to deficiencies or toxicities under the soil conditions of their new environment.

In other cases, adaptive alleles may have deleterious pleiotropic effects. Malaria adaptation can be conferred by the HbS variant of the *HBB* gene; when heterozygous this allele gives a ten-fold reduction in risk of severe malaria, but results in sickle cell anaemia when homozygous (Hill et al. 1991; Archer et al. 2018). Other adaptations to malaria via different genes also often result in deleterious blood-related disorders, such as G6PD deficiency, $\propto^{+}$ thalassemia and hemoglobin C (Kwiatkowski 2005). Otherwise, adaptations to African sleeping sickness, cold ambient temperature and low amino acid levels have been associated with higher risk or prevalence of chronic kidney disease, migraine and celiac disease, respectively (Genovese et al. 2010; Wang et al. 2012; Mathieson et al. 2015a; Key et al. 2018).

There is also some evidence of population-specific adaptation that results in differences in the outcome of treatment of non-inherited disorders. For example, a derived variant of *IFNL4,* which when homozygous results in a loss of the IFN-λ4 protein, is inferred to have evolved under positive selection in Eurasian populations and has been associated with a more rapid clearing of the hepatitis C virus infection (Key et al. 2014). The African Biaka population also appear to have multiple alleles (*CUL5, TRIM5, APOBEC3G*) fixed or at high frequency that appear to confer protection to HIV (Zhao et al. 2012).

Hence, studies into local adaptation are not only imperative to understanding the selected, and by extension neutral, proportion of the modern human genome, but are vital in understanding which populations are most at risk from modern disease or malnutrition. Moreover, such studies help to identify adapted alleles that are functionally relevant to critical environmental pressures. In turn, this may allow a deeper understanding of the genetic basis of human phenotypes, including those relating to disease or disease risk.

1.4.3. A Note on Diversity in Modern Human Studies

Local genetic differences can contribute to population differences in the genetic basis of common disease, as well as to response to treatment. Therefore, a comprehensive understanding of disease progression and treatment in individual populations cannot be achieved without a fundamental understanding of the underlying genetic diversity of such a population, including that which may have been driven by local adaption events. However, it is clear that genetic studies have been historically biased towards European populations, especially those exploring genetic associations with disease (Sirugo et al. 2019). This bias not only results in a failure to capture the full extent of global genetic diversity (Popejoy and Fullerton 2016), but also exacerbates global health inequalities by limiting our knowledge of health-related traits to well-studied populations (Sirugo et al. 2019).

When population-specific mutations drive disease, a biased understanding of the genetic drivers of disease to only well-studied populations have been shown to increase health and diagnosis disparities between populations. This is the case for retinitis pigmentosa (where over 3000 mutations in 65 genes have been identified in causing retinitis pigmentosa, but mostly in Europeans (Sirugo et al. 2019)) and cystic fibrosis (where the most common causal variants differ between European and African populations (Padoa et al. 1999; Stewart and Pepper 2017)). Moreover, causal mutations unique to an under-studied population may remain unidentified and the associated disease underdiagnosed or untreated, as was the case for the mutation driving transthyretin amyloid cardiomyopathy in African Americans, an underdiagnosed cause of heart failure (Buxbaum et al. 2006; Sirugo et al. 2019).

More than this, the bias towards certain portions of genetic diversity in human genomic studies reduces our ability to accurately, and therefore safely, translate genetic research into clinical care of under-studied populations. Genome-wide association studies (GWAS) have most commonly been undertaken in European populations (52% of studies in European populations as of 2018 (see **Figure 1.4;** (Sirugo et al. 2019)). This has resulted in estimates of the genetic risk of variants in Europeans, but it is unclear to what degree the genetic determinants of certain health-related traits are shared by other populations (Huang et al. 2022). Differences in genetic architecture (as a result of drift or local selection events in populations of different ancestry (Lim et al. 2014)) as well as differences in linkage disequilibrium (which can affect the accuracy of identifying causal variants and varies according to demographic history (Tishkoff et al. 2009)) both heavily contribute to the lack of replication of GWAS-estimated risk values amongst populations.

Using estimates of risk calculated from one population in another population of differing ancestry may result in an inaccurate estimation of clinical risk, delayed or lack of intervention and could falsely prioritise different treatment or drug strategies,

ultimately most adversely affecting under-studied populations (Sirugo et al. 2019). This, partnered with the clear bias of precision medicine towards well-studied, particularly European, populations should emphasise the need for genetic studies across a range of historically under-represented populations, as well as a recognition of the current health inequalities facing many populations today.

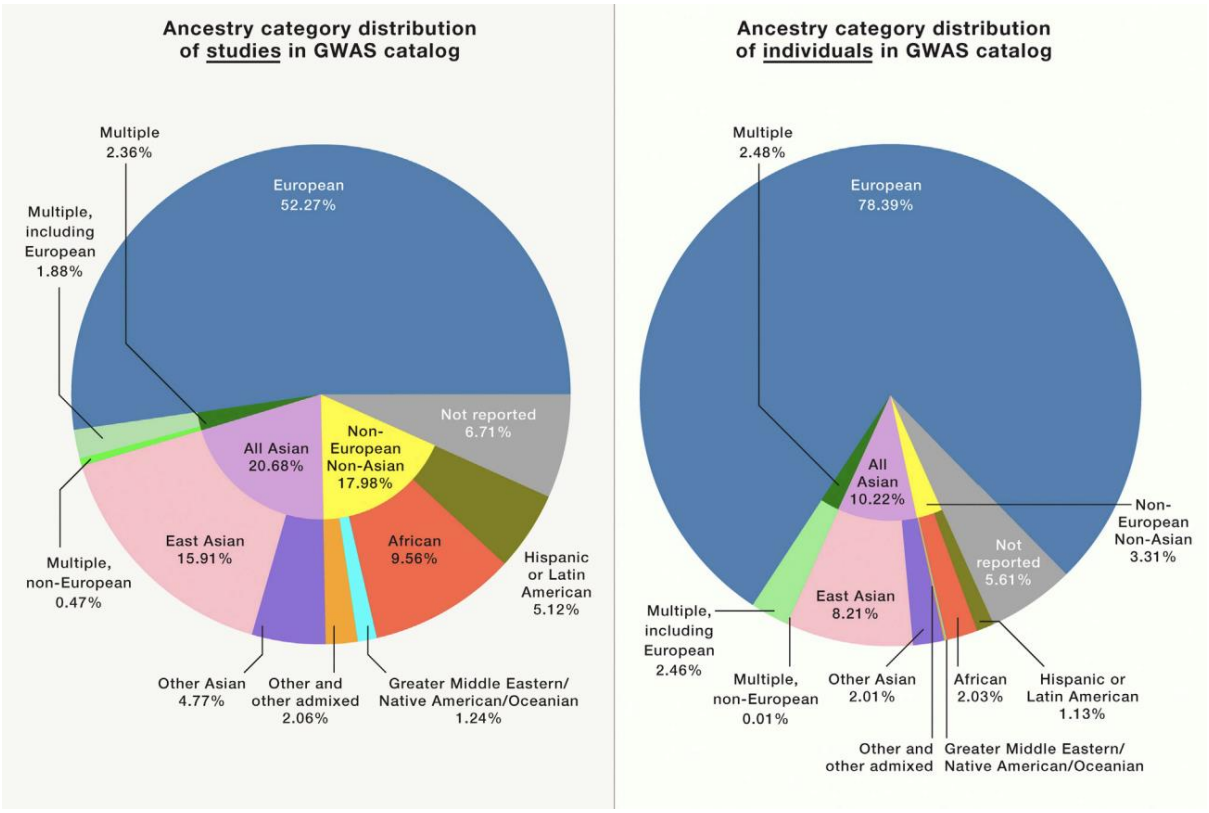


***Figure 1.4*: *Summary of GWAS by Ancestry*.** *A summary of the ancestry by GWAS (left) and by individuals within each GWAS (right), as calculated by the GWAS Catalogue through January 2019. Taken from (Sirugo et al. 2019).*

1.5. Genomics of Local Adaptation

1.5.1. Dynamics of Local Adaptation

Positive selection driving human local adaptation occurs under highly variable and complex scenarios. Not only is this selection exerted on different populations (or groups of populations), it also acts on various different phenotypes, at different times, and at various strengths. More than this, positive selection to drive an adaptive trait can be exerted on alleles of different origins, and may be either monogenic or polygenic in nature, of which a summary is given below.

1.5.1.1. Origin of Selected Allele(s)

The origins of beneficial alleles can be broadly split into those coming from *de novo* mutation, from standing variation (or previous polymorphisms within a population) or introduced into a population via admixture or introgression (Hermisson and Pennings 2005; Peter et al. 2012; Rees et al. 2020). Local adaptation in humans has been suggested to be mediated from alleles of all three origins, with each form of selection resulting in subtly different signatures on the genome (see **Figure 1.5;** (Rees et al. 2020).

Selection on *de novo* mutation (SDN) acts on a new allele that is immediately advantageous in its environment, and therefore rapidly increases in frequency in the population if selection is strong (termed a hard selective sweep (Pritchard et al. 2010; Rees et al. 2020). Still, there are limits to its occurrence. The adaptive mutation must not only appear in a population experiencing an at least somewhat unmediated selective pressure, but also avoid immediate loss from the population due to random genetic drift. Variants that act as dominants, therefore having an effect in heterozygotes, can immediately be under strong selection and are less likely to be quickly lost from a population (Rees et al. 2020).

Selection on standing variation (SSV) differs from SDN in many senses, including the appearance of the mutation with respect to the onset of selection. Here, previously segregating alleles become advantageous following a change in selective pressure, often when encountering a new environment (via change of the current environment or migrations into novel conditions; (Hermisson and Pennings 2005, 2017; Peter et al. 2012)). Hence, the selected allele is older than the selective pressure. Such an allele may have been previously maintained in the population by neutral processes (being neutral or nearly neutral, see **Section 1.2.1**) or maintained by balancing selection (Andrés 2011; Bitarello et al. 2018). The latter has great potential in contributing alleles for rapid adaptation in novel conditions, since they by definition affect phenotype and fitness, and are already maintained at intermediate frequencies (Andrés 2011; de Filippo et al. 2016; Bitarello et al. 2018).

SSV is considered likely very important in mediating human local adaptation, specifically in populations rapidly encountering novel environments following the Out of Africa migration (Hermisson and Pennings 2017; Rees et al. 2020). The sudden onset of selective pressure in these populations would have given little time for the emergence of *de novo* mutations (especially considering the small effective population size contributing to a low effect mutation rate), and the rapid population growth allows many low-frequency polymorphisms with selective potential to be maintained in the population (de Filippo et al. 2016; Hermisson and Pennings 2017). Indeed, a number of recent studies have suggested that SSV has been prevalent, if not dominant (Hernandez et al. 2011; Pybus et al. 2015), in human adaptation (Peter et al. 2012; Schrider and Kern 2017).

Finally, admixture between genetically distinct populations can introduce beneficial alleles into a population, a process termed adaptive admixture. This has been implied many times between modern human populations, particularly when gene flow between populations accompanies cultural exchange (such as alleles conferring lactase persistence spreading from pastoralists to close-by populations, recorded between both African and European populations (Tishkoff et al. 2007a; Evershed et al. 2022). Other

examples include the contribution of the Duffy blood group allele conferring malaria resistance from an ancient African population to modern Malagasy populations (Pierron et al. 2018).

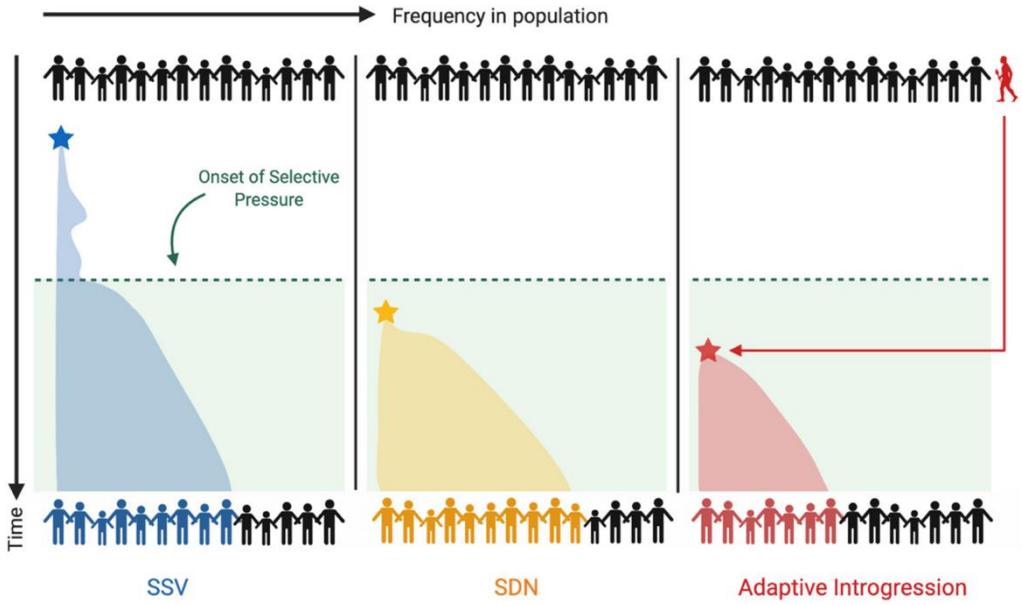


***Figure 1.5**: **A schematic representation of the rise of allele frequency through a population according to its origin**. Stars represent mutations (blue: variant present in the population prior to selection; yellow: mutation occurring after the onset of selection; red: mutation present in an archaic population which spreads through a receiving population following an admixture event). The frequency of each variant in a population following a selection event is represented by the number of people icons of their respective colours under each panel. The red walking person icon in the top right represents an archaic human, which contributes the red variant to a receiving population following an admixture event. Taken from (Rees et al. 2020).*

Adaptive introgression, or the adaptive admixture between modern and archaic humans, has also been identified as playing a role in mediating local selective pressures in non-African populations. Since Neanderthal and Denisovans long inhabited Eurasia before modern humans (a short time after the divergence between archaic and modern humans; approximately 600,000 years ago (Schaefer et al. 2021)), these archaic populations had time to develop their own local adaptations, for which different modern humans could rapidly acquire through various introgression events when first encountering these novel environments (Reich et al. 2010; Prüfer et al. 2014; Rees et al. 2020).

Indeed, whilst many of the non-neutral introgressed alleles were deleterious and therefore removed by purifying selection in modern humans (Juric et al. 2016), a few others have been shown to contribute to adaptations related to immune function, pigmentation and oxidative stress (see **Figure 1.6**). They have been highlighted in their

role of conferring virus resistance, particularly against RNA viruses in Europeans (Enard and Petrov 2018). In some instances, where genetic diversity in of itself is advantageous, more than one archaic allele is maintained (Dannemann et al. 2016; Enard and Petrov 2018).

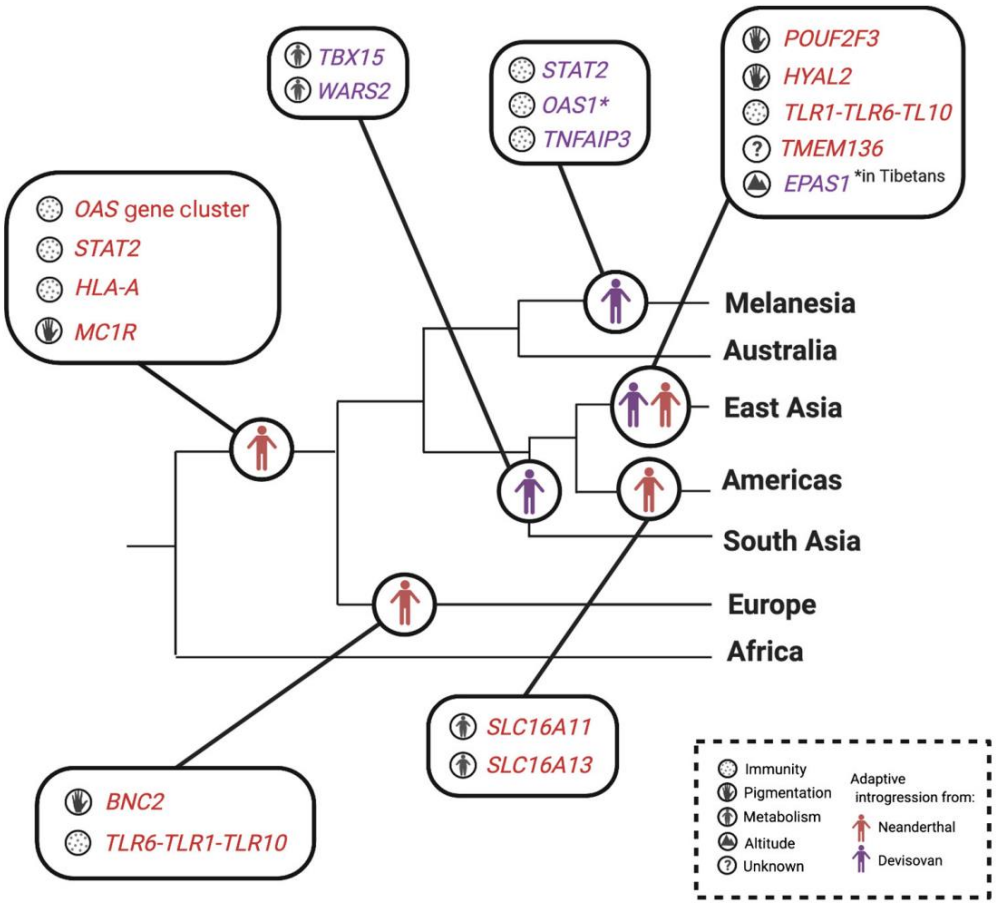


***Figure 1.6*: Cartoon representation of the adaptive introgression events from archaic human populations into modern humans.** *The red and purple persons indicate human populations with evidence of adaptive introgression with Neanderthals and Denisovans, respectively. The gene(s) believed to be under selection given in the linked boxes, on the lineages where selection is suggested to have occurred. Taken from (Rees et al. 2020).*

1.5.1.2. Polygenicity of Adaptation

Selection may be mediated by one beneficial allele (monogenic selection) or many beneficial alleles (polygenic selection) (Pritchard et al. 2010). Monogenic selection can result in strong selection signatures in a single locus, and therefore are often the simplest signatures to identify in the genome, with most known adaptations monogenic in nature (*e.g.,* (Perry et al. 2007; Tishkoff et al. 2007a; Minster et al. 2016; Ilardo et al. 2018; Key et al. 2018; Pierron et al. 2018; Evershed et al. 2022)).

Polygenic selection is characterised by selection acting on multiple adaptive alleles in a population. This does not necessarily include any one extreme single allele frequency change, but rather a group of alleles that all show concerted shifts in allele frequencies to shift the phenotype in the adaptive direction (Pritchard et al. 2010). Which alleles mediate adaptation depends on stochastic processes and their pleiotropic restraints, which in turn are specific to their particular genetic background (and therefore also influenced by epistasis; (Phillips 2008; Solovieff et al. 2013)). Hence, different populations may show different alleles responding to selection, or at different degrees.

Polygenic adaptation has been proposed to be common in modern humans, since most traits are likely polygenic, with many alleles mediating phenotypic response. Indeed, various studies have suggested polygenic selection is prevalent throughout human history, and it has been proposed to drive adaptations to diet, metabolism, pathogen resistance and altitude (Fumagalli et al. 2011; Daub et al. 2013; Berg and Coop 2014; Daub et al. 2015; White et al. 2015; Nédélec et al. 2016; Gouy et al. 2017, 2017; Roca-Umbert et al. 2022).

1.5.1.3. Epigenetic Local Adaptation

Whilst not directly relevant to the work in this thesis, it should also be stated here that there is increasing interest in the importance of epigenetics in local adaptation.

Epigenetic response is a somatically heritable change of chemical modification, most commonly studied being DNA methylation, that does not result in changes in the DNA sequence. Whilst it is still debated if such chemical modifications are heritable (Heard and Martienssen 2014), it has been shown that epigenetic responses occur under changes in the local environment, particularly during development (Gokhman et al. 2017). These modifications can occur much faster than genetic adaptations, and so it has been proposed that they also may help populations to mediate environmental pressures over periods as short as a lifetime (Gokhman et al. 2017).

Whilst is it difficult to show that epigenetic changes are adaptive rather than a response to stress, adaptation and epigenetic change has been linked in populations of Central Africa. Here, genetic variants associated with methylation variation have been identified to show signatures of positive selection (Fagny et al. 2015), leading to the suggestion that epigenetic change allows rapid, plastic mediation of selective pressures before adaptive alleles can be cemented in the genome.

1.5.2. Signatures of Local Adaptation

Selection events leave distinct patterns, or signatures, in the genome. These signatures rely heavily on not only the strength and timing of selection, but also the allele origin. Large allele frequency differentiation in a SNP between populations, more extreme than could be explained by neutral demographic processes, is considered an almost-universal signature of strong local adaptation (Rees et al. 2020). However, linked variation, which can be a powerful tool to identify loci under selection, is highly varied according to an allele's selective history.

Alleles that rapidly rise in frequency, as under SDN, exhibit linked haplotypes of low diversity and many low frequency variants (see **Figure 1.7**). A selective sweep is defined by this extended haplotype homozygosity surrounding the selected site, accompanied by high population differentiation and skews in the site frequency spectrum (or an excess of high-frequency derived alleles; (Sabeti et al. 2006; Pritchard

et al. 2010)). Such a rapid rise in frequency also leaves a distinctive pattern on the allele's genealogy; a long internal branch with short terminal tips (or a "star"-like shape) represents the sweep of an advantageous allele through a population (Field et al. 2016). The strength of these signatures is largely dependent on the strength and timing of selection, with those most striking signatures pertaining to site frequency spectrum, population differentiation and haplotype homozygosity originating from strong selection events with an onset surrounding < 80kya, 75-50kya and <30kya, respectively (Sabeti et al. 2006).

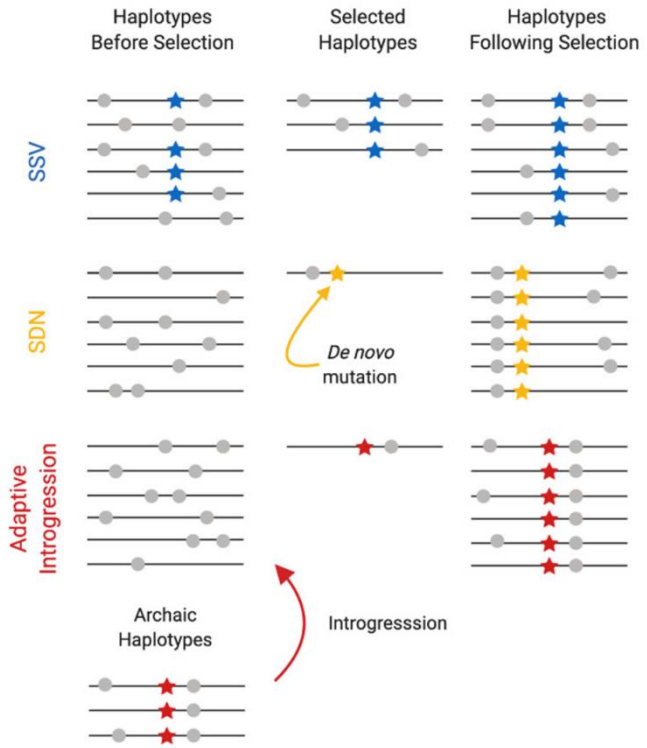


***Figure 1.7**: **Cartoon depicting the haplotypes arising from selection on standing variation (SSV), selection on de novo mutation (SDN), and adaptive introgression**. Stars represent the beneficial allele (blue: mutation occurs in the population prior to selection; yellow: mutation occurring after the onset of selection; red: mutation occurs in an archaic population which spreads through a receiving population following an admixture event) and grey circles represent neutral linked polymorphic alleles. Taken from (Rees et al. 2020).*

However, under SSV, the signatures of linked variation are usually highly reduced and less easily distinguishable from the neutral genetic background (**Fig. 1.7**). This is not only due to a typically less striking rise in frequency (if the selected allele is already segregating at intermediate frequencies), but the selection occurring on multiple haplotype backgrounds (Rees et al. 2020). It is this lack of haplotype homozygosity that particularly separates the signatures of SSV from SDN. SSV also encompasses a range of ratios between allele origin and selection onset (*e.g.,* selection acting on an allele swiftly following its origin compared to selection acting on an allele more than, say, 80,000 years following its origin). This highly varied age and frequency of the segregating allele further increases the variance of expected linked variation under the umbrella term of SSV. In some cases, SSV may even resemble, or be considered as, SDN if the onset of selection closely follows the allele origin or a population bottleneck results in extreme frequency increase (Wilson et al. 2014).

The selection signatures arising from adaptive admixture events differ heavily from those of SDN or SSV, and have a dual aspect to their identification. Admixed or introgressed segments must be first extracted from shared ancestry or incomplete lineage sorting (Huerta-Sánchez et al. 2014), before then isolating adaptive segments from those that are neutral. These admixed or introgressed segments are largely defined by the timescale and populations of the gene flow event. For example, introgressed segments are long and unusually similar to archaic segments but appear young, are accompanied by high levels of linkage disequilibrium and only present in some modern human populations (Yang, Malaspinas, et al. 2012; Liang and Nielsen 2014; Racimo et al. 2015).

Many of the classic signatures used to identify positive selection are therefore also present in neutrally introgressed segments, such as long-range LD or population differentiation (Racimo et al. 2015). The strongest evidence for adaptive introgression is thus usually considered an unusually high frequency of an introgressed segment compared to the empirical distribution of introgressed segments throughout the genome. Similarly, strong evidence for recent adaptive admixture, that between modern human populations, is often identified by higher proportions of putatively adaptive ancestry compared to the expected ratio of ancestry components, as well as by clustering algorithms which classify individuals by their genetic patterns (such as *STRUCTURE* or *ADMIXTURE* (Pritchard et al. 2000; Alexander et al. 2009; Pierron et al. 2018; Wangkumhang and Hellenthal 2018; Secolin et al. 2019). Naturally, identifying such signatures rely on confident assignment of ancestry components, be that from modern or archaic humans, and may be biased by limited ancient DNA data of ancestral human populations.

Finally, local adaptation may be inferred from signatures of polygenic adaptation, which can be summarised as highly varied and often weak signatures spread over many loci (see **Figure 1.8**). Multiple small frequency shifts, which may occur at different timepoints and can be highly spread throughout the genome, means that polygenic selection can appear indistinguishable from neutral genetic drift (Pritchard et al. 2010; Le Corre and Kremer 2012). The highly variable dynamics of polygenic selection, such as the number of loci under polygenic adaptation, as well as the effect sizes and the origin of these alleles, also result in selection signatures that may appear very different from each other. In some cases, even very strong signatures may accompany polygenic adaptation; alleles with large effect sizes or a small number of alleles without deleterious pleiotropic effects may sweep to fixation as expected under SDN (Chevin and Hospital 2008).

The omnigenic model offers an additional explanation as to why polygenic selection is difficult to identify in modern humans. This model considers that variants across almost the entire genome can contribute to an adaptive phenotype; these variants are found in both “core” genes (those which directly affect the phenotype) and “peripheral” genes (those that indirectly affect the phenotype through interacting networks (Boyle et al. 2017; Mathieson 2021)). The effects of these “core” genes are consistent over different populations and are therefore more likely to be identified as under positive selection. However, the effects of “peripheral” genes are governed by their interaction with a number of other peripheral and core genes. Hence, their effects may differ amongst studied populations according to the differing allele frequencies across the entire gene network, and they may constitute the more variable signatures that elude identification (Mathieson 2021).

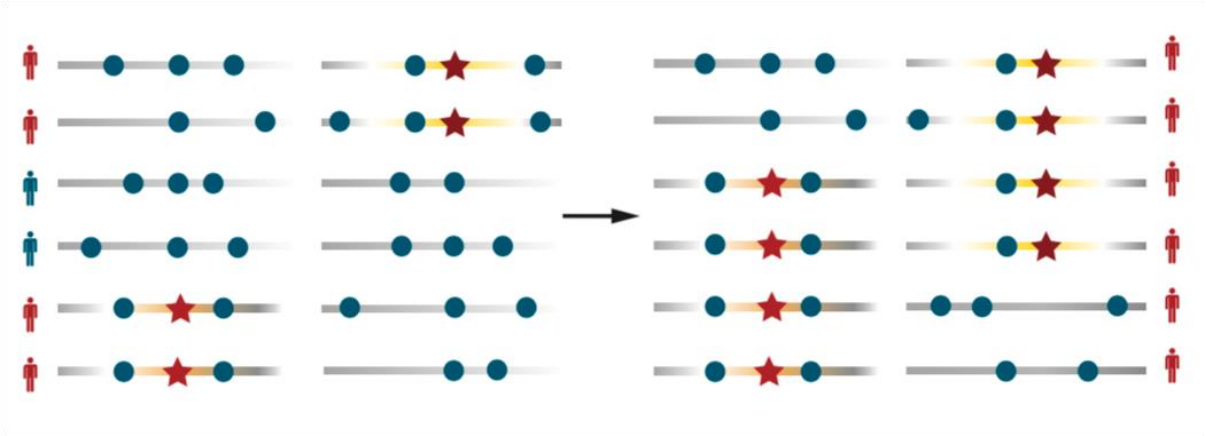


***Figure 1.8**: **Cartoon depiction of polygenic adaptation**. Each horizontal bar represents a haplotype and stars indicate mutations occurring on different genetic backgrounds. Taken from (Fan et al. 2016).*

1.5.2.1. A Note on Hard and Soft Sweeps

Historically, selection signatures have been categorised into either "hard" or "soft" sweeps (Hermisson and Pennings 2005; Peter et al. 2012; Schrider and Kern 2016, 2017), where hard sweeps represent the rapid rise of frequency resulting from SDN and soft sweeps represent weaker, and altogether more variable, signatures on the genome. These soft sweeps may be a result of weak selection, SSV or recurrent mutations, although the latter is unlikely in humans due to the low effective mutation rate (Hermisson and Pennings 2005). Many studies have attempted to determine the relative importance of hard and soft sweeps in human adaptation, more recently using methods like Approximate Bayesian Computation or Supervised Machine Learning to compare the proposed incidence of SDN, SSV and polygenic selection (Peter et al. 2012; Schrider and Kern 2016, 2017).

Whilst there is value in assigning categories to differing dynamics of selection, the dynamics themselves (*e.g.,* strength, age and frequency of the selected allele) are so intrinsically heterogeneous that a discrete categorisation will never fully represent signatures on the genome (Rees et al. 2020). Whilst many have moved away from using terms such as "hard" or "soft" sweeps (or at least recognise that "soft" can often be simply interpreted as the definition of all sweeps that aren't strong enough to be "hard") and instead moved to using terms such as SDN and SSV (Peter et al. 2012), often these terms are still used and may falsely imply a binary, or less variable, nature of selection in modern humans.

1.6. Methods to Identify Local Adaptation

Local adaptation can be represented in the genome via a multitude of selection signatures. Methods, either individually or as a collection within the field, must be able to identify these various signatures and, by extension, identify the different dynamics of local adaptation. This includes different strengths and timing of selection, different allele origins, and different numbers of alleles under selection, all which must be distinguished from neutral processes.

Random genetic drift can increase the frequency of alleles, sometimes so rapidly that they can mimic the allele frequency rise seen under positive selection. Demographic processes, particularly population bottlenecks, can aid this frequency increase (Rees et al. 2020). Such bottlenecks are not only common in human populations, but many populations have only a partially resolved demographic model (Gravel et al. 2011, 2013). Without a high degree of certainty surrounding demographic history, it can be difficult to tease apart past population expansions and admixture with that of positive selection (Pritchard et al. 2010; Peter et al. 2012).

Purifying selection can also mimic some of the genomic signatures of local adaptation. Background selection, or the reduced effective population size at sites linked to those undergoing purifying selection, increases the effect of genetic drift (Charlesworth et al. 1995). This can cause strong genetic differentiation between populations, which is usually an indicator of local adaptation (Cruickshank and Hahn 2014). There is also evidence that deleterious introgressed alleles result in heterosis, or hybrid vigour, in modern humans, again mimicking signatures of adaptation (Kim et al. 2018; Zhang et al. 2020).

1.6.1. Summary Statistics

The most common method to identify signatures of local adaptation is to represent selection signatures (or particular aspects of the overall signature) as a summary statistic and compare these to a neutral background. In practice, this involves sampling many loci throughout the genome, calculating the chosen summary statistic and identifying the loci with summary statistic values unexpected under neutrality (Sabeti et al. 2006; Rees et al. 2020). The gold-standard approach would be to have a neutral expectation based on truly accurate neutral simulations, operating under a fully resolved demographic model. In reality, this is both highly improbable and impractical for many studies and populations. A common method is instead to build an empirical distribution of the summary statistic values throughout the genome and identify loci with outlier values. It is an important distinction here that those outliers do not necessarily have signatures unexpected under neutrality, but are strong candidate targets of selection compared to the rest of the genome (Rees et al. 2020).

Many classical summary statistics used to identify local adaptation aim to identify one of three signatures of selection (see **Figure 1.9**): high-frequency derived alleles (or skews in the site frequency spectrum (Tajima 1989)), population differentiation (Weir and Cockerham 1984; Yi et al. 2010; Yassin et al. 2016; Crawford et al. 2017; Librado and Orlando 2018; Schmidt et al. 2019) and degree of haplotype homozygosity (Voight et al. 2006; Sabeti et al. 2007; Ferrer-Admetlla et al. 2014; Szpiech et al. 2021). Since these aspects of selection signatures are all strong under strong SDN, many classic summary statistics are well-equipped to identify this type of selection (indeed, reflected in the

literature of known adaptations (Rees et al. 2020)), but comparatively poorly equipped at detecting weaker signatures of selection (with some exceptions *e.g.,* (Ferrer-Admetlla et al. 2014; Garud et al. 2015; Field et al. 2016; Speidel et al. 2019)).

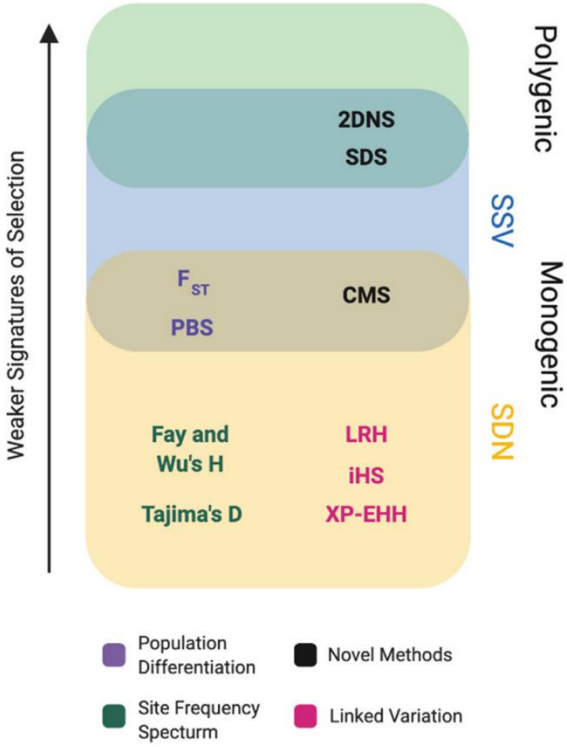


***Figure 1.9**: **Common statistics used to identify local adaptation in modern humans.** Statistics (not a comprehensive list) arranged according to their power in identifying different strengths and modes of local adaptation signatures. Abbreviations: iHS, integrated haplotype score; LRH, long range haplotype; PBS, population branch statistic; SDS, singleton density score; XP-EHH, cross-population extended haplotype homozygosity. Taken from (Rees et al. 2020).*

As previously noted, strong population differentiation can be considered a more universal signature of adaptation, leading to a particular focus on allele frequency differentiation methods to identify potentially weaker signatures of selection. This includes F_{ST} and similarly operating methods, such as the population branch statistic (PBS; comparing three pairwise F_{ST} values between three populations to identify unusual differentiation (Yi et al. 2010)) and its derivatives (PBS_{n1}, PBS_{nj} and PBE (Yassin et al. 2016; Crawford et al. 2017; Schmidt et al. 2019)). Multiple methods (*e.g.,* $Bayenv$ (Günther and Coop 2013)) continue to add value to using population differentiation methods to more confidently identify SNPs which are unusually differentiated by estimating the covariance due to shared ancestry in allele frequencies between populations.

There has been some development in using composite statistics, or those that combine multiple summary statistics whilst accounting for correlation between the individual statistics, to present an overall score for selection (Grossman et al. 2010; Ma et al. 2015). Naturally, the composite score is reduced when any of the individual summary statistics used are themselves of low value. Hence, these methods have the highest power in identifying those selective events that result in consistently strong selection signatures. More than this, combining signatures into a single score in this way can

make understanding the dynamics of selection less intuitive. Calculating such statistics individually but manually considering their results together for a candidate locus may result in a more informed view on the probability and nature of selection.

1.6.1.1. Machine Learning Methods

Machine learning methods integrate spatial patterns along the genome to classify loci into prespecified models of evolution, such as the exact nature of inferred selective sweeps (including traditional classifications of "hard" or "soft", as well as completeness or timing of the sweep; Pybus et al. 2015; Sheehan and Song 2016; Schrider and Kern 2016, 2017; Kern and Schrider 2018; Sanchez et al. 2020; Gower et al. 2021, Caldas et al, 2022; Qin et al. 2022). In some cases, the spatial patterns of a range of summary statistics are used to represent the genomic data (Pybus et al. 2015; Sheehan and Song 2016; Schrider and Kern 2016, 2017; Sugden et al. 2018; Mughal and DeGiorgio 2018). In the case of deep learning algorithms, a particularly promising subset of machine learning algorithms, a predefined set of summary statistics is not required as input (LeCun et al, 2015; Sheehan and Song 2016; Kern and Schrider 2018; Sanchez et al, 2020; Gower et al, 2021, Caldas et al, 2022, Qin et al. 2022). Instead, deep learning methods can effectively use the entirety of the available raw data to learn which features are most useful for predicting the nature and presence of natural selection, potentially improving inferences by using the data which would be lost in the process of calculating summary statistics (Korfmann et al, 2023).

Machine learning methods in their entirety are particularly promising since they are trained by a range of simulations modelling different selection scenarios, some of which, *e.g.,* those arising from polygenic adaptation or SSV, are not represented by extreme patterns in the raw data, including those captured by summary statistics (Pybus et al. 2015; Sheehan and Song 2016; Schrider and Kern 2016, 2017; Kern and Schrider 2018; Sanchez et al, 2020; Gower et al, 2021, Caldas et al, 2022, Qin et al. 2022). In short, these methods have the potential to recognise the more subtle genomic signatures that characterise other strengths or dynamics of selection.

1.6.2. Ancient DNA

Ancient DNA (aDNA) sequencing methods have vastly improved in the 21st century alone, with available ancient human samples now sequenced in their thousands (Racimo et al. 2015; Wohns et al. 2022). Whilst many of these are dated to more recent times (*e.g.,* <5000 years ago), there is increasing coverage of ancient human populations as far back as ~45kya (Skoglund and Mathieson 2018), as well as numerous archaic human samples dating as more than ~50kya (Green et al. 2010; Reich et al. 2010; Meyer et al. 2012, 2012; Castellano et al. 2014; Prüfer et al. 2014, 2017; Mafessoni et al. 2020).

Ancient DNA is exceptionally powerful in selection studies because it can provide direct snap-shots of past allele frequencies (Key et al. 2018). This allows the identification of rapid allele frequency change, and hence potentially suggest the onset of selection; identify or support proposed selection at individual sites; and allow a deeper understanding of the role of adaptation and neutrality in creating modern population differentiation. (Sverrisdóttir et al. 2014; Mathieson et al. 2015a; Mathieson and Mathieson 2018; Le et al. 2022; Mathieson and Terhorst 2022).

More than this, aDNA is imperative in understanding adaptive admixture or introgression, where ancient genomes can identify the population origin of alleles and

identify any regions with an unusually high contribution from one ancestral population, that which is expected under adaptive admixture (Mathieson et al. 2015a; Racimo et al. 2015). Ultimately, aDNA has played a key role in evaluating the role of gene flow in neutral and selected genetic diversity, *e.g.,* has allowed identification of loci mediating adaptation post-admixture in European and American populations (Mathieson et al. 2015a; Lindo et al. 2016), as well as introgressed alleles conferring adaptation in many non-African populations (Green et al. 2010; Reich et al. 2010; Meyer et al. 2012; Huerta-Sánchez et al. 2014; Prüfer et al. 2014; Racimo et al. 2015).

1.6.3. Tree-based Methods

Many recent advances have been made in genealogical reconstruction methods (Hejase et al. 2020): those that build individual trees for SNPs along the genome which, in theory, represent an almost complete history of each locus. Here, the evidence for selection can then be more directly evaluated from the inferred tree. This is considered superior to using classical summary statistics to infer selection, since these statistics complex evolutionary or genomic patterns into a single value (Rees and Andrés 2022).

Tree-based methods to infer selection have been used in a number of programmes for some time (*e.g., ARGweaver* and *msprime*; both continuing to be developed (Rasmussen, Hubisz, et al. 2014; Kelleher et al. 2016; Hubisz and Siepel 2020; Baumdicker et al. 2022, Brandt et al, 2022) but only recently have computational advances allowed the inference of genealogies for large sample sizes (Kelleher et al. 2019; Speidel et al. 2019; Wohns et al. 2022), pushing these methods to the forefront of the field. Two notable recent methods, *Relate* and *tsinfer*, are able to efficiently build trees from over 1000 samples, including the integration of high-coverage archaic human samples (Kelleher et al. 2019; Speidel et al. 2019; Wohns et al. 2022). *tsinfer* has also been used to infer a "unified genealogy", one that has been built from 3601 modern and high coverage ancient genomes (alongside using 3589 low-coverage ancient samples to further constrain and date the tree) and represents the most complete tree-representation of the genetic history of humans yet (Rees and Andrés 2022; Wohns et al. 2022).

Whilst *tsinfer* is accompanied by a python package (*tskit* (Kelleher et al. 2019; Baumdicker et al. 2022)) that allows manipulation and analysis of the inferred tree sequences, including calculating summary statistics using the inferred trees, *Relate* has a built-in method for inferring selection, and is generally considered more suited to this analysis (Speidel et al. 2019). *Relate* first infers a tree for each SNP along the genome (see **Chapter 2**), and then simultaneously re-estimates branch lengths, changing population size through time and mutation rate to refine the tree sequences. Its selection test uses these inferred trees and allele ages to compare the spread of a mutation (or the lineage carrying a mutation) to all other lineages, conditioning on the number of lineages present on the onset of the mutation and outputting a probability of the mutation's trajectory under neutrality. This has shown to be successful in identifying both monogenic and polygenic adaptation (Speidel et al. 2019), and its direct inference of a mutation's spread is suggested to be better suited in identifying subtler instances of selection.

1.6.4. Environmental Correlations

Arguably the strongest evidence of local adaptation is when correlations are observed between allele frequencies and environmental factors, given that they are more extreme than expected under population histories and relatedness (Günther and Coop 2013). *Bayenv* is one such method that is able to account for population structure when testing for environmental correlations, and has been used to identify human adaptations along clines of climate, diet and pathogen density (Hancock et al. 2008; Hancock, Witonsky, et al. 2010, 2011; Fumagalli et al. 2011; Hancock, Clark, et al. 2011).

Otherwise, linear models may be used to ask to what extent shared ancestry and proposed environmental factors explain the observed allele frequencies, a method that has been used to infer adaptation to climate, including that of the cold receptor *TRPM8* to cold ambient temperature (Raj et al. 2013; Key et al. 2018). Evidence for local adaptation is also given when functionally relevant genes are inferred to be under selection in a linked extreme environment (as the case for strong signatures of selection in *AS3MT*, a gene associated with arsenic metabolism, in populations living on arsenic-rich soils of Argentina (Schlebusch et al. 2015)), or when several populations experiencing similar selective pressures show signatures of selection in the same gene(s) (as is the case for selection signatures surrounding the *LCT* gene, that which is responsible for lactase persistence in multiple pastoralist populations (Tishkoff et al. 2007a; Gerbault et al. 2011; Evershed et al. 2022)).

1.6.5. Identifying Polygenic Selection

Identifying polygenic adaptation is considered far more challenging than identifying monogenic adaptation, since the signatures of selection are defined by weaker allele frequency changes spread amongst multiple loci (as well as being highly variable in the degree of allele frequency change, number of loci and their associated effect sizes). The most common methods used to identify polygenic selection are derived from genome-wide association studies (GWAS) (Sabeti et al. 2006; Berg and Coop 2014; Berg, Zhang, et al. 2019) (see **Section 1.6.5.1**), alongside a growing use of gene set and gene network methods (Daub et al. 2013, 2017; Gouy et al. 2017; Gouy and Excoffier 2020) (see **Section 1.6.5.2.**).

Additional methods to identify polygenic adaptation, but not explored in detail here, included statistics such as $2DNS$ (a McDonald-Kreitman-based test (Daub et al. 2015)) or SDS (Single Density Score; which uses the distances in tree tip branch lengths to infer selection (Field et al. 2016)). The latter statistic, whilst shown to have good power in identifying recent selection, is still liable to population stratification when using GWAS hits (see **Section 1.6.5.1**).

1.6.5.1. Genome-Wide Association Studies

The increasing availability of genome sequences and catalogues of human genetic variation makes GWAS a popular choice in identifying polygenic selection (Sabeti et al. 2006). These studies scan for genetic markers over complete genomes that are associated with a particular trait or disease, and their ultimate objective is to identify causal genetic variants and estimate their corresponding effect size on such a trait. To identify polygenic adaptation, the effect sizes calculated from these studies can also be

used to identify the global sets of alleles that show positive covariance (Berg and Coop 2014; Berg, Zhang, et al. 2019).

However, it has now been suggested that differences in the genetic basis of traits in populations and hidden population stratification amongst the samples used in these studies can result in spurious claims of polygenic adaptation (Berg, Harpak, et al. 2019; Sohail et al. 2019). Still, population stratification is only believed to result in small, systematic biases, rather than false genome-wide significant associations. Hence, when testing if derived mutations increasing or decreasing a phenotype are enriched for evidence of positive selection, one study chose to use only SNPs with genome-wide significant associations in their analysis (Speidel et al. 2019). This method, which also only used effect direction rather than effect size, reduces confounding due to population stratification, but does not completely avoid it (Speidel et al. 2019).

However, it must also be considered that SNPs identified as causal to a trait in one population may not necessarily be causal in other populations, and can further lead to false inferences of selection. Many of the recognised issues with using GWAS highlight the need for wider sampling of human populations in order to understand currently undocumented genetic variance, and its association with modern traits (see **Section 1.4.3**).

1.6.5.2. Gene Set Methods

Methods using gene sets, or gene networks, propose to combine the weak signatures from multiple genes within a meaningful set, such as known biological pathways (Daub et al. 2013, 2017; Foll et al. 2014; Amorim et al. 2015; Gouy et al. 2017; Gouy and Excoffier 2020). This can result in a gain of statistical power to detect polygenic selection, even when the selection on individual genes is weak. These approaches have given evidence for polygenic adaptation to pathogens (Daub et al. 2013), convergent adaptation to high-altitude (Foll et al. 2014) and tropical forest environments in modern humans (Amorim et al. 2015), as well as more ancient selection acting after the human-chimp split (Daub et al. 2017).

One particular gene-set enrichment test, notable for its simplicity and customisation, is the $SUMSTAT$ method (Daub et al. 2013). This method uses the sums of test statistics, compared to neutral gene sets, to detect selection for a given gene set or pathway. Hence, it specifically looks for the signatures of small effect mutations over a phenotype, making it highly suitable for identifying polygenic selection. Moreover, whilst this method has previously used F_{ST} as the summed test statistic (Daub et al. 2013), $SUMSTAT$ can integrate any test statistic in its identification process, allowing the use of more sensitive statistics should they be identified or proposed.

Still, $SUMSTAT$ is naturally limited by its integrated test statistic, and may falsely identify signatures arising from background selection or relaxation of constraint as positive polygenic selection (Daub et al. 2013). As well as this, such a method relies on accurate sets of neutral genes, which in practice are often sets of random genes throughout the genome that should, but do not necessarily, approximate neutrality.

All gene set or network methods also inherently rely on functionally related sets or networks of genes, where some genes may be associated with multiple functions and hence represented in multiple sets or networks (Daub et al. 2013; Gouy et al. 2017). It can therefore be difficult to tease apart selection on one function from another and,

given pleiotropic constraints, it is often unrealistic to expect all genes within a gene set to be exhibiting signatures of polygenic selection.

1.6.6. Determining the Selective Driver

Once local adaptation has been identified, the natural next step is to identify the driving selective force. In some cases, this is integrated in the identification of local adaptation itself (*e.g.,* when using environmental data to identify unusual allele correlations, repeated adaptation of the same genes to similar environments or functional relationships between putatively selected genes and environment, see **Section 1.6.4**), but in most cases additional avenues may be used.

Determining the timing of proposed local adaptation events is a common way of attempting to identify a putative selective force. Here, aDNA is especially valuable, since it can help reconstruct the allele frequency through time, constraining estimates of the onset of selective pressures (Mathieson et al. 2015a; Mathieson and Mathieson 2018; Le et al. 2022; Mathieson and Terhorst 2022). Due to the increasing availability of samples, many studies using aDNA to infer selection have been carried out in European populations. Their results have questioned the proposed link between the development of agriculture and selection inferred on the *FADS* locus (linked to fatty acid metabolism), *AMY1* (production of amylase) and *LCT* (production of lactase) (Sverrisdóttir et al. 2014; Mathieson et al. 2015a; Mathieson and Mathieson 2018; Le et al. 2022).

Otherwise, studies may aim to establish the function of putatively selected genes or genomic elements to suggest the environmental pressure for which they are responding, often using model organisms (Lamason et al. 2005; Fujimoto et al. 2008). Such methods have helped elucidate the role of genes such as *SLC24A5* (under positive selection in European populations; affecting pigmentation in zebrafish (Lamason et al. 2005)) and *EDAR* (under positive selection in Asian populations; affects mammary and eccrine glands in mice (Fujimoto et al. 2008) and now shown to affect hair thickness, tooth and ear shape, sweat gland density and chin protrusion in modern humans (Fujimoto et al, 2008; Adhikari et al, 2015; Reyes-Reali *et al*, 2018, Kataoka et al, 2021).

Determining the phenotypic trait affected by human variants is considerably more challenging. Whilst large association studies can propose trait associations (with recognition of the bias towards more studied and sampled populations), elucidating the molecular response of selected variants relies on further analyses integrating transcriptomics, metabolic and microbiota datasets (Rees et al. 2020). SNP-function may also be categorised with the use of high-throughput assays exploring the effect of proposed adaptive variants on protein expression, transcription or methylation (Downes et al. 2019). As stem cell technology improves, it may also become more commonplace to use pluripotent stem cells and stem cell-derived organoids to experimentally test the phenotypic consequences of certain gene variants in human cells (Kilpinen et al. 2017; Hwang et al. 2019). This represents an exciting future avenue for local adaptation studies, providing additional tools to investigate not only variants under selection, but the functional consequences of such selection, ultimately necessary to validate these signatures.

1.7. Micronutrients in the Human Diet

Micronutrients are an essential part of the human diet since, with the exception of vitamin D, they are not synthesised in the body (Shenkin 2006). Instead, they must be

consumed via the diet, where the levels of micronutrients in plant and animal foodstuffs rely heavily on the underlying soil geology (Diamond 2002; White et al. 2015; Dhaliwal et al. 2019). They play a central role in human metabolism and maintenance of tissue function, and are particularly important in immunity and healthy growth and development (Shenkin 2006). Micronutrient levels in the diet outside a very small but specific healthy range can result in a range of pathologies, many of which are very common over global or individual populations (see **Section 1.7.1**).

Micronutrients themselves can be split into two main categories: vitamins (organic compounds made by plant and animal sources) and minerals (inorganic compounds absorbed from soil or water; (Tako 2019)). Minerals are further subset into macrominerals and trace minerals, where macrominerals are needed in slightly higher levels compared to trace minerals and vitamins (see **Table 1.2**), but are still required at far reduced levels compared to macronutrients such as carbohydrates or fats (Prasad 2013; Tako 2019). Here, we focus on the trace minerals and macrominerals in the human diet.

***Table 1.2**: **Recommended daily amounts (RDA; for adults >19 years old) for all micronutrients needed for maintaining human health**. Given alongside major sources of each micronutrient. Taken from (Streit 2018; Rowles 2023).*

Micronutrient Type	Micronutrient	Source	RDA
Macrominerals	Potassium	Lentils, acorn squash, bananas	4700mg
	Sodium	Salt, processed foods	2300mg
	Calcium	Milk products, leafy greens, broccoli	2000-2500mg
	Chloride	Seaweed, salt, celery	1800-2300mg
	Phosphorus	Salmon, yogurt, turkey	700mg
	Magnesium	Almonds, cashews, black beans	310-420mg
	Sulphur	Garlic, onions, brussels sprouts, eggs, mineral water	None established
Trace Minerals	Iron	Oysters, white beans, spinach	8-18mg
	Zinc	Oysters, crab, chickpeas	8-11mg
	Fluoride	Fruit juice, water, crab	3-4mg
	Manganese	Pineapple, pecans, peanuts	1.8-2.3mg
	Molybdenum	Beans, lentils, grains, organ meats	2000mcg
	Copper	Liver, crabs, cashews	900mcg
	Iodine	Seaweed, cod, yogurt	150mcg
	Selenium	Brazil nuts, sardines, ham	55mcg
Vitamins	Vitamin A	Liver, dairy, fish, sweet potatoes, carrots, spinach	700-900mcg
	Vitamin B1 (thiamine)	Whole grains, meat, fish	1.1-1.2mg
	Vitamin B2 (riboflavin)	Organ meats, eggs, milk	1.1-1.3mg

Vitamin B3 (niacin)	Meat, salmon, leafy greens, beans	14-16mg
Vitamin B5 (pantothenic acid)	Organ meats, mushrooms, tuna, avocado	5mg
Vitamin B6 (pyridoxine)	Fish, milk, carrots, potatoes	1.3mg
Vitamin B7 (biotin)	Eggs, almonds, spinach, sweet potatoes	30mcg
Vitamin B9 (folate)	Beef, liver, black eyed peas, spinach, asparagus	400mcg
Vitamin B12 (cobalamin)	Clams, fish, meat	2.4mcg
Vitamin C (ascorbic acid)	Citrus fruits, bell peppers, Brussels sprouts	75-90mg
Vitamin D	Sunlight, fish oil, milk	600-800 IU
Vitamin E	Sunflower seeds, wheat germ, almonds	15mg
Vitamin K	Leafy greens, soybeans, pumpkin	90-120mcg

1.7.1. Micronutrient Deficiency and Toxicity

Micronutrient deficiency is estimated to affect 2 billion people worldwide, with the majority of these individuals in sub-Saharan Africa and South-Central Asia (Bhutta and Salam 2012; Bailey et al. 2015). Of these, 178 million are children under 5 and estimated to have experienced stunted growth from micronutrient deficiency, with 19 million of these predicted to be at such a level of malnutrition to be at a risk of death (Bhutta and Salam 2012). Often deficiencies co-occur, and may be further coupled with protein or caloric malnutrition (Bailey et al. 2015). This can complicate the association between micronutrient deficiency and health, since it most explicitly associates undernutrition to increased health risk.

Still, micronutrient deficiencies independently result in an increased risk of metabolic, cardiovascular and infectious diseases (Shenkin 2006; Triggiani et al. 2009; Tulchinsky 2010; Bailey et al. 2015; Biban and Lichiardopol 2017; Khan et al. 2022). During development, deficiencies may result in stunted growth, mental retardation and an overall increased risk of morbidity and mortality (Halsted et al. 1972; Yant et al. 2003; Conrad et al. 2004; Shenkin 2006; Prasad 2013; Bailey et al. 2015). Hence, pregnant women and children under five are considered the most vulnerable to the long-term effects of micronutrient deficiency and generally are the focus of public health intervention strategies. In some cases, intervention strategies have great success and health disorders may even be reversed with supplementation of the missing micronutrients. In other cases, particularly when malnutrition occurs at key periods of development, the health consequences remain permanent (Bailey et al. 2015).

The most widespread macromineral and trace minerals deficiencies are those pertaining to iron, iodine and zinc[2] (Bhutta and Salam 2012; Bailey et al. 2015). Iron is the most common global deficiency, with approximately 40% of children between 6-59 months and 36% of pregnant women estimated as anaemic in 2019 (Stevens et al. 2022). Anaemia increases the risk of poor maternal and perinatal health, delays growth and cognitive development, and considerably reduces physical work capacity and impairs immune and endocrine function (Stevens et al. 2022). Goitre, the swelling of the thyroid gland as a result of iodine deficiency, is observed in approximately 15.8% of the global population (Gebremichael et al. 2020). Like iron, extreme iodine deficiency is tightly associated with impaired cognitive function and mental retardation, particularly during development. Zinc deficiency, estimated to affect 1.1billion people worldwide, however, is primarily associated to impaired immune function (Bailey et al. 2015; Khan et al. 2022). Deficiency is associated with increased risk of diarrhoea and acute respiratory infections, including the SARS-CoV-2 virus, which are major causes of death in many global populations (Khan et al. 2022). The major health consequences and symptoms associated with less common macromineral and trace mineral deficiencies of interest, alongside their toxicity symptoms, are summarised in **Table 1.3.**

Micronutrient toxicities generally result in increased gastrointestinal distress, nausea, vomiting and diarrhoea, with some claims that they can increase the risk of poisoning from non-essential minerals (Peraza et al. 1998; Pike and Zlotkin 2019). Toxicities have been identified across many domestic animal and plant species, often a result of the underlying soil conditions (Becker and Asch 2005; Giri et al. 2021; Kaur and Garg 2021), but are less commonly recorded in humans compared to micronutrient deficiencies (Fraga 2005). This may be due in some parts to the decoupling of toxicity from surrounding malnutrition risk. Hemochromatosis, or the systemic overload of iron caused by mutations in the *HFE, HAMP, HJV, TFR2* and *SLC40A1* genes, is the most common micronutrient toxicity disorder (Brissot et al. 2018). This is most common in European populations (see **Section 1.7.3.**) and results in a range of symptoms including chronic fatigue, joint pain and, in extreme cases, cardiac failure (Naugler 2008; Brissot et al. 2018).

[2] Deficiencies of vitamin A and folate are also very common micronutrient deficiencies and are of a concern to global health. Vitamin A deficiency is the leading global cause of vision loss (Xu et al. 2021)) and deficiencies of folate, or vitamin B9, is estimated to be associated with 80% of neural tube defects during pregnancy (fatal or severely disabling birth defects that result in approximately 300,000 cases worldwide (Wald 2022)).

***Table 1.3**: **The role of thirteen essential trace minerals and macrominerals.** Given alongside documented symptoms or diseases associated with their respective deficiencies and toxicities.*

Micronutrient	Role	Deficiency Symptoms	Toxicity Symptoms	References
Potassium	Nerve transmission, muscle function	Increased blood pressure, fatigue, constipation, polyuria, cardiac arrhythmias	Neuromuscular dysfunctions	(Erdman et al. 2012; Jain et al. 2013; Stone et al. 2016; Streit 2018)
Sodium	Maintains blood pressure	Impaired cognition, fatigue, nausea, weight loss	Increased blood pressure, hypertension, cardiovascular morbidity	(Geerling and Loewy 2008; Hurley and Johnson 2015; Grillo et al. 2019)
Calcium	Bone and teeth structure and growth, muscle function, blood vessel contraction	Reduced bone strength (osteoporosis), defective bone mineralisation and bone softening (osteomalacia) rickets (in children)	Weight loss, polyuria, heart arrhythmias, fatigue, soft tissue calcifications	(Sunyecz 2008; Calcium et al. 2011; Streit 2018)
Chloride	Maintains fluid balance, digestive juices	Muscle weakness, lethargy, loss of appetite	Pulmonary irritation and injury (gaseous explosure)	(Grossman et al. 1980; Morim and Guldner 2022)
Phosphorus	Forms bone and cell membrane structure	Anaemia, muscle weakness, bone pain, osteomalacia, decreased immunity	Hypotension, vascular calcification, cardiac arrest	(Razzaque 2011; Streit 2018)
Magnesium	Enzymatic reactions, regulates blood pressure	Nausea, vomiting, fatigue, weakness, seizures, muscle cramps, hypocalcemia, hypokalemia, osteoporosis	Hypotension, nausea, muscle weakness	(Castiglioni et al. 2013; Al Alawi et al. 2018; Streit 2018; Ajib and Childress 2022)
Iron	Supplies muscles with oxygen, hormone synthesis	Anaemia (fatigue, shortness of breath, dizziness, heart palpitations)	Hormonal abnormalities, decreased immunity, diabetes, heart disease, liver disease, fatigue, joint pain	(Fraga and Oteiza 2002; Fraga 2005; Streit 2018; Stevens et al. 2022)
Zinc	Growth, immunity	Delayed growth, impaired immune function, alopecia, diarrhoea, cognitive decline	Anaemia, headache, abdominal cramps, nausea	(Plum et al. 2010; Streit 2018; Khan et al. 2022)
Manganese	Carbohydrate, amino acid and cholesterol metabolism	Abnormal bone and cartilage development,	Neurological dysfunction	(Horning et al. 2015; O'Neal and Zheng 2015; Streit 2018)

		delayed wound healing		
Molybdenum	Cofactor for enzyme reactions	Tachycardia, night blindness, irritability, childhood death (if a result of the genetic disorder molybdenum cofactor deficiency)	Hallucinations, seizures, cognitive decline	(Novotny 2011; Reiss and Hahnewald 2011; Rowles 2023)
Copper	Brain and nervous system function, connective tissues	Anaemia, ataxia, low numbers of white blood cells (neutropenia)	Vomiting, abdominal pain, paralysis	(Williams 1983; Ashish et al. 2013; Prohaska 2014)
Iodine	Thyroid regulation	Growth and development impairment, neurodevelopmental deficits, cretinism, hypothyroidism and goitre	Nausea, diarrhoea, vomiting, delirium	(Miles 1998; Biban and Lichiardopol 2017; Streit 2018; Southern and Jwayyed 2022)
Selenium	Thyroid regulation, reproductive health, defence against oxidative damage, potential cancer prevention	Cognitive decline, impaired immunity, osteoarthritis (*e.g.,* Kashin-Beck disease), cardiomyopathy (*e.g.,* Keshan diseasd), exacerbate iodine deficiency	Metallic taste in mouth, hair and nail loss, nausea, diarrhoea, fatigue, nervous system abnormalities	(MacFarquhar et al. 2010; Streit 2018; Ibrahim et al. 2019; Shi et al. 2021; Xu et al. 2022)

1.7.2. Global Variation of Micronutrient Levels

A significant proportion of contemporary micronutrient deficiency is linked to the socioeconomic status of individual populations, and significantly associated with global poverty and undernutrition (Keats et al. 2019). Indeed, the most prevalent cases of micronutrient deficiency are observed in low-income and middle-income countries (Keats et al. 2019; Khan et al. 2022). Often, the diets of these countries largely consist of staple foods that do not sufficiently cover the range of nutrition needed for optimum health, and may be very low in levels of specific micronutrients (Ishfaq et al. 2021). Toxicities, however, often result from chemical exposure to the individual, rather than at a population-wide level (Fraga 2005).

Still, the public health concerns of some populations include particular micronutrient deficiencies, or even toxicities, driven by the micronutrient levels of their underlying soil environment. This can result in micronutrient-associated diseases endemic to a population or region. Here, such a health burden on a population may have been experienced for considerably longer periods of time, rather than a product of relatively recent global and societal inequality.

1.7.2.1. Soil Geology and Micronutrient Levels

Micronutrients are present in soils in a variety of different forms, which vary in their bioavailability (the extent that they may absorbed and used) to plant and animal species. They may be present as precipitates, adsorbed onto soil particles, present as complex ring compounds or simply be part of rocks (primary minerals) or clays (secondary minerals) (Keefer 1999). Many chemical factors affect the form micronutrients take in the soil and their corresponding bioavailability, such as other available elements, pH and organic matter (Dhaliwal et al. 2019; Liu et al. 2021). Geographical factors also play a role, particularly the surrounding rock types, topography and distance from the ocean (*e.g.,* coastal regions are noted as particularly high in iodine and many iodine deficient regions are highly landlocked (Cifor 2006; Shetaya et al. 2012)). It is important to note that high levels of a micronutrient in a soil does not directly result in high intake in the human diet, as it is the bioavailability itself that plays the greatest role.

Given the extreme variety of global environments, soils can be highly variable even between relatively proximal localities (and also may not align directly with modern country or region classifications). Whilst there are few comprehensive studies of micronutrient levels across global soils, and even fewer that compare the levels of different micronutrients, the global distributions of some micronutrients are well-explored. Often, this is linked to either the prevalence of human micronutrient-associated disorders (*e.g.,* the endemic diseases caused by selenium deficiency in East Asia have prompted many studies investigating selenium levels in local soils (Hurst et al. 2013; Liu et al. 2021, p. 202)), or their relevance in agriculture, particularly to optimise plant growth (Diamond 2002; Alloway 2013; Duborská et al. 2022).

Areas of the world with the most notable, and most well resolved, deficient or toxic soils for different micronutrients are given below, accompanied by their associated endemic diseases when relevant. These are simply well-documented examples, and should not be considered the only examples of micronutrient deficiency or toxicity in global soils. Moreover, there is likely considerable change from the ancestral soil state following the birth of wide-scale agriculture, and contemporary micronutrient soil levels may not reflect their levels throughout the majority of time (Diamond 2002).

Selenium

Selenium levels have been shown to be highly variable at both the global and the local scale. Most notably, selenium-deficient soils have been recorded in areas of East Asia, particularly along a wide belt stretching across the southwest to northeast of China (although this also contains pockets of selenium-enriched soils; see **Figure 1.10** (Xia et al. 2005; Liu et al. 2021)). Indeed, endemic diseases related to selenium deficiency have been identified in particularly rural populations within these regions, such as the cardiomyopathy Keshan disease and bone disorder Kashin-Beck disease (Shi et al. 2021; Xu et al. 2022). Otherwise, there is evidence for geospatial variation in selenium levels across many African soils, with deficiencies particularly highlighted in Malawi (Hurst et al. 2013; Ligowe et al. 2020), alongside New Zealand, Finland, Australia and some areas of North and South America (**Fig 1.10;** (Koivistoinen and Huttunen 1986; Thomson 2004; Jones et al. 2017).

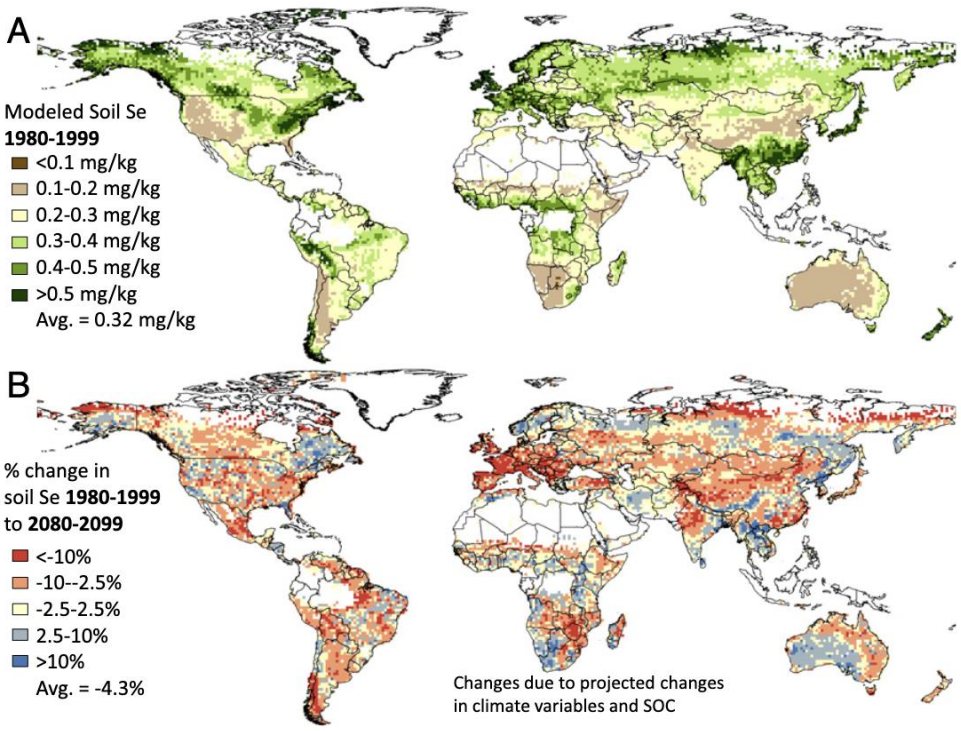


***Figure 1.10*: Geographical representation of soil selenium levels.** *Soil selenium levels modelled from 1980-1999 (A) and predicted percentage change in soil selenium levels from 1980-1999 to 2080-2099 (B). Taken from (Jones et al. 2017).*

Iodine

Iodine deficiency often co-occurs with selenium deficiency in soils, due to their shared reliance on proximity to aquatic environments (Winkel et al. 2015; Duborská et al. 2022). Soils documented as low in both selenium and iodine include those in Central Africa, Central and East Asia and pockets in the Americas, amongst others (Lyons 2018). This often results in the co-occurrence of selenium and iodine-associated disorders. Moreover, the metabolic pathways which rely on selenium and iodine are often tightly interlinked (Duborská et al., 2022), where low selenium levels may even exacerbate the effects of iodine deficiency (Triggiani et al. 2009).

Rainforest environments have been particularly highlighted as being low in iodine, including those in Central Africa, the wet zones of Sri Lanka and the wet, monsoon delta regions of Java and Bali (Cifor 2006). Goitre has been reported at high incidence in some populations living in these environments, including the Bantu population of Central Africa, which have a 42.9% incidence of goitre (Dormitzer et al. 1989). Other affected populations include those in Central and South America, with the incidence of goitre at 54.6% in Mexico in the 1980s (Hetzel and Nutrition 1988), and various South and East Asian populations, where goitre has been treated with iodine supplementation since the mid 19th century (Miles 1998).

Zinc

Zinc deficient soils have been identified particularly across the Middle-East (Sillanpaeae 1982; Ryan et al. 2013), as well as India and sub-Saharan Africa (Arunachalam et al. 2013; Kihara et al. 2020), and some areas of China, Indonesia and north-western region of South America (Prasad 2013). From a survey of over 3500 soils across 29 countries, Iraq was found to have the highest proportion of zinc-deficient soils (57%) followed by Turkey (35%) and Pakistan (20%) (Sillanpaeae 1982). Indeed, zinc deficiency also has the strongest history in the Middle-East, where the first instances of zinc deficiency were recognised in the 20th century (Halsted et al. 1972; Gibson 2012; Prasad 2013). Here, the dietary zinc levels were so low that it resulted in extreme stunted growth, delayed sexual development and recurrent infections that usually resulted in death before 25 years of age (Halsted et al. 1972; Prasad 2013; Khan et al. 2022).

Sodium and Chloride (Salt)

Hyper-salinity, or excess of salt in soils (associated with levels of both sodium and chloride, the elemental constituents of salt) often occur in arid zones with low rainfall, and has been linked to the fall of many agricultural civilisations when it has decimated crop growth (including multiple times in the history of Iraq (Shahid et al. 2018)). Hence, it must be noted that the main recorded impact excess salinity has on human health is the reduction of crop yield and general nutrition, rather than overt micronutrient deficiencies or toxicities. As well as in Iraq (and other areas of the Middle-East), hyper-salinity has been reported in the arid regions of South Africa, the Americas and Australia (Nell and van Huyssteen 2018; Shahid et al. 2018; Hassani et al. 2021), but it is unclear the degree to which recent agriculture has contributed to the contemporary excess levels of salt in soils (Hassani et al. 2021).

Phosphorus

Phosphorus levels in the soil are also heavily affected by farming practices, both by use of fertilizers or by over-farming (Dhaliwal et al. 2019; Alewell et al. 2020). Still, calcareous soils are known to have low bioavailability of phosphorus (von Wandruszka 2006), as well as low levels of iron (Chen and Barak 1982). There is also a broad pattern of increased soil phosphorus in non-African soils, particularly across northern areas of Europe, Asia and the Americas (He et al. 2021).

Micronutrients with Limited Soil Data

The soil levels of the remaining micronutrients of interest (see **Table 1.3**) are less clearly elucidated. From the literature, we highlight extremely high levels of magnesium in some areas of Central Asia (Vyshpolsky et al. 2008; Karimov et al. 2009); low levels of potassium in Ethiopia and New Zealand (Edmeades et al. 2010; Laekemariam et al. 2018) and potassium-rich soils in India (Naidu et al. 2011); calcium deficiency of the coastal plain of the south-eastern United States (Adams and Hathcock 1984); low levels of copper amongst peat soils such as those in Japan, South Africa, Scandinavia and Russia, amongst others (Alloway and Tills 1984); high levels of molybdenum in sedimentary based soil but low levels in acidic soils (Barceloux and Barceloux 1999); and toxic levels of manganese in Puerto-Rico, Brazil, areas of tropical Africa and eastern Australia (Fernando and Lynch 2015).

1.7.3. Adaptation to Dietary Micronutrients

Micronutrient levels in the diet are strong candidates for local adaptation in modern humans for two key reasons. The first is that they are necessary for maintaining optimum health and development, but with complete reliance on what is absorbed via the diet (with the exception of vitamin D). Secondly, micronutrient levels are highly variable across different soil environments, therefore exerting potentially strong differential selective pressures over modern human populations. This proposed local adaptation may also be polygenic in nature, given the many genes associated with the transport and uptake of different dietary components (including micronutrients (Monteiro et al. 2015)).

Below, a summary of the studies suggesting local adaptation in response to micronutrient levels is given (see **Figure 1.11**), proposed to be driven by either underlying soil levels or cultural factors.

Iodine

Strong signatures of selection, identified using a modified version of the *lnRsb* method which searches for unusual haplotype homozygosity amongst populations (Tang et al. 2007), in the iodide-dependent thyroid pathways have been inferred in two African pygmy populations, both of which live on iodine-deficient rainforest soil environments (Herráez et al. 2009). Since changes to the thyroid hormone have also been shown to result in short stature, this has been used to suggest that the short-stature of different populations across the world may be a phenotypic consequence of their adaptation to their iodine-deficient tropical forest environments.

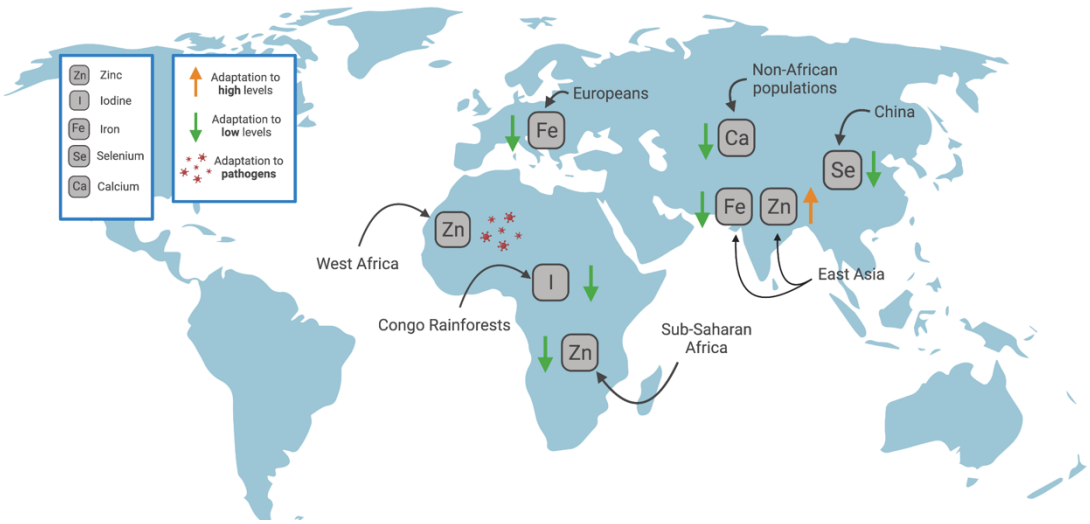


Figure 1.11: Schematic map of suggested local adaptation events in modern humans in response to micronutrients*. Proposed examples of micronutrient-associated adaptation in modern humans alongside their suggested driver* (Distante et al. 2004; Hughes et al. 2008; Herráez et al. 2009; Engelken et al. 2014; White et al. 2015; Ye et al. 2015; Zhang et al. 2015a; Engelken et al. 2016; Roca-Umbert et al. 2022)*. Includes the instance of selection on HFE in European populations, which has been argued to be a false positive owing to allele surfing (Peischl et al. 2016). Made by biorender.com.*

Selenium

White *et al* (2015) suggested that selenium-associated genes (selenoproteins and those that regulate selenium or selenocysteine, see **Section 1.8.1**) show evidence of positive selection in the populations living in regions of the world documented with low selenium soil levels. These genes were enriched for signatures of differentiation (as calculated via F_{ST} (Weir and Cockerham 1984)) in populations of Central South Asia and East Asia (White et al. 2015). Within this latter region, which has a particularly high prevalence of extreme selenium deficiency and associated disorders (Xia et al. 2005; White et al. 2015; Shi et al. 2021; Xu et al. 2022), the enrichment of F_{ST} signatures were localised to the populations living on soils of low selenium levels, particularly the Hezhen, Naxi and Oroqen populations of China (White et al. 2015). This study suggested a polygenic nature of adaptation to selenium levels, which is supported by an additional study inferring signatures of selective sweeps across three selenium-associated genes in East Asians (*GPX1, GPX3, SELENBP1* (Engelken et al. 2016).

Zinc

Zinc adaptation has also been suggested to be polygenic in nature (Zhang et al. 2015a; Roca-Umbert et al. 2022). Zinc concentration is regulated in the body by a family of 24 zinc transporters, with this entire gene set inferred to show an unusual degree of differentiation between Eurasian and African populations (Zhang et al. 2015a; Engelken et al. 2016; Roca-Umbert et al. 2022). Some zinc transporters also show especially strong evidence of positive selection. This includes *SLC30A9*, which has been inferred to be under selection to regulate zinc levels, but in opposite directions, in East Asians and Africans (Zhang et al. 2015a). A correlation was shown between the haplotype under selection and the zinc levels in soil or crops, suggestive of positive selection in response to the low and high levels of zinc in the diets of Africans and East Asians, respectively.

Other zinc transporters with notable evidence of positive selection include *SLC39A4,* which appears to be differentiated between West Africans and Eurasians at a level that is inconsistent with coalescent simulations of neutrality (Engelken et al. 2014). Here, the African variant of *SLC39A4* has been suggested to reduce zinc uptake and consequent availability in the human body, thereby starving pathogens of zinc. Thus, suggested as an adaptive response to the pathogen-rich environment of sub-Saharan Africa (Engelken et al. 2014; Zhang et al. 2015a). This "pathogen-starvation" hypothesis has not only been suggested in playing a role in zinc regulation, but also notably in the regulation of iron, amongst other key micronutrients needed for pathogen development (Pietrangelo 2015).

Iron

Dietary changes in recent human history, particularly those that resulted from the agricultural revolution approximately 10,000 years ago (Naugler 2008; Brown et al. 2009; Latham 2013), have been suggested to have driven putative adaptation in iron-associated genes. The early agricultural diet was largely characterised by staple crops and had reduced nutritional variety, as well as severe reductions of micronutrients such as iron and calcium (Diamond 2002; Naugler 2008). Indeed, it has been suggested that the high frequency of the *C282Y* allele of the *HFE* gene in Europeans, which results in hemochromatosis or excess iron levels, is a result of adaptation to dietary iron deficiency that followed this cultural change (Distante et al. 2004; Naugler 2008). Parallel adaptation of this gene has also been suggested in East Asian populations,

driven by traditional, low iron diets, rather than underlying soil levels (Ye et al. 2015). Others have suggested the putative adaptation of *HFE,* in European populations, is a response to the colder European climates, since iron plays a key role in thermoregulation (Distante et al. 2004). In more recent years, however, there is growing evidence that the high frequency of the *C282Y* allele in Northern European populations is a result of allele surfing on waves of range expansions from South-East to North-West Europe, rather than as a result of positive selection (Peischl et al. 2016).

Calcium

Suggested adaptation in calcium-associated genes has also been associated with recent changes in the human diet, rather than underlying soil levels. This is the tentatively suggested selective driver of the putative parallel adaptation in the *TRPV6* gene in non-African populations, as inferred by extended haplotype homozygosity, but with little supporting evidence (Hughes et al. 2008). Alternatively, since vitamin D is required to absorb calcium from the diet, and vitamin D synthesis in turn depends on UV exposure, it has also been suggested that lower UV levels may drive adaptations to increase calcium absorption. This gains support from the correlation between signatures of positive selection in calcium-associated genes with the latitude of northern European populations (Mathieson and Terhorst 2022).

1.7.3.1. Public Health Connotations

Global soil micronutrient levels are changing as a result of climate change, rising CO_2 levels and over-farming (see **Fig. 1.10;** (Shahid et al. 2018; Dhaliwal et al. 2019; Hassani et al. 2021)). This, alongside increased migration and mobility of global populations, means that many populations will likely encounter micronutrient levels for which they lack adaptations, or even have adaptations to regulate in the opposite and now deleterious direction (the "evolutionary mismatch" scenario (Manus 2018). It is therefore a matter of global health to understand how varying micronutrient levels, especially deficiencies and toxicities, may interact with different genetic backgrounds.

Many public health policies have benefited from an understanding of the adaptive history and modern phenotypic consequences of populations. For example, it is now UK Public Health policy to strongly recommend those of self-identified African and South Asian descent to take vitamin D supplements in their diet, more so than those of European descent (http://www.gov.uk). This stems from the recognition that lighter skin pigmentation is an adaptation to decreased UV levels, allowing the body to absorb more UV and maintain vitamin D synthesis (Carlberg 2022). Those of darker skin pigmentation, but who live in environments where UV levels are lower, are therefore more susceptible to decreased UV absorption and deficient vitamin D levels.

It is important to note that such policies operate at a population level, and an understanding of a population's adaptive history offers an understanding of health risk at only the population level, rather than for each individual. Historically, there have been issues with conflating this subtle, but key, distinction. For example, sickle cell disease has long been considered a "black disease" given its prevalence in West African populations (a result of the causal allele conferring malaria resistance when heterozygous (Esoh and Wonkam 2021)). This has resulted in many instances of misdiagnosis, where general symptoms experienced by those of individuals of African ancestry have been falsely attributed to sickle-cell, as well as diagnoses of sickle cell unconsidered in those of non-African ancestry (Yudell et al. 2016). Similarly, cystic

fibrosis is underdiagnosed in those of African ancestry due to its reputation of a "white disease" (Yudell et al. 2016). Ultimately, this results in delayed medical treatment and significant emotional and physical distress of the individual, and serves as a warning of using population level generalisations at the individual level.

1.8. Micronutrients in Wider Biology

Much of the research into micronutrient biology, outside of human health, has been done in the frame of agricultural science (Welch and Graham 2005; Shukla et al. 2009; Singh 2009; Bouis and Welch 2010). High and healthy crop yields rely on the correct proportions of micronutrients, particularly manganese, molybdenum, nickel, zinc and iron, as well as arsenic, cadmium, lead and tin potentially playing an essential role at lower concentrations (Alloway 2013). Animal farming also relies on optimum levels of copper, manganese, molybdenum, zinc and iron, as well as chromium, cobalt, selenium and vanadium (Alloway 2013). Whilst the role of micronutrients in agriculture is outside the scope of this thesis, it is worth noting the key role that research on understanding the micronutrient conditions of global soils and biofortification will play in meeting the increased demands of a growing human population and addressing global health inequalities (Tulchinsky 2010; Dhaliwal et al. 2019; Hassani et al. 2021).

Still, the acquisition, absorption and digestion of these key dietary components have affected many aspects of organism evolution, some of which have been reviewed here: (McWilliams 2011; Swanson et al. 2016; Xu et al. 2021). When considering the role micronutrients play in adaptive evolution across species, is important to note that the exact levels required of each micronutrient, or even what is classified as a micronutrient, may vary over divergent taxa. For example, in plants, phosphorus is considered a macronutrient since it contributes a significant amount of energy and resources for plant growth, but in humans is considered a micronutrient since it is needed in much smaller quantities and is involved in more specific metabolic processes (Alloway 2013). Therefore, the adaptive response and compensatory mechanisms of these taxa facing phosphorus deficiency or toxicity can be expected to substantially vary.

The micronutrients which are essential in the diet versus those that can be synthesised by the organism also differs amongst taxa; some species are able to synthesise some micronutrients within the body that other taxa may be forced to consume via the diet, increasing their reliance on, for example, local soils or foodstuffs. A notable example of this is the changing reliance on dietary vitamin C across vertebrates. Taxa such as teleost fishes, anthropoid primates (the group that includes humans) as well as some bat, rodent and bird species have lost the ability to synthesis vitamin C *in vitro* owing to mutations in *GLO* (Cui et al. 2011; Drouin et al. 2011). Many hypotheses exist for why this gene has been pseudogenised across these taxa, the most relevant to human evolutionary history being that the increased availability of ascorbate-rich fruit in the diet of ancestral anthropoid primate ancestors rendered *in vitro* synthesis superfluous (Hornung and Biesalski 2019).

Despite there being biological necessity for most human-classified micronutrients across wider organisms, the range of functional roles and differential reliance on the uptake of different micronutrients makes it difficult to extract specific evolutionary trends across broad groups of taxa. Still, selenium in wider vertebrate evolution, specifically in respect to its catalytic role in selenoproteins, is well elucidated. Here, we

give an overview of selenoprotein evolution and provide this as an additional example to explore how micronutrients may affect genome evolution across non-human species.

1.8.1. Selenoprotein Evolution

Selenium is an essential micronutrient for vertebrates, with an especially narrow range over which it is nutritionally optimal (see **Table 1.2;** (Sarangi et al. 2017)). Selenium levels above or below this range result in deficiencies and toxicities across vertebrates, as reported in humans and many agricultural species. For example, in humans, mild deficiencies can result in reduced immune function, lower fertility and cognitive decline, with extreme deficiencies, as identified in some areas of China, resulting in diseases of the heart and bone (Shi et al. 2021; Xu et al. 2022). In ruminants, white muscle disease is associated with extreme selenium deficiency, with less extreme deficiencies leading to reduced fertility and incidence of mastitis and metritis (Spears and Weiss 2008; Hefnawy and Tórtora-Pérez 2010; Sordillo 2013). Farmed animals have also been shown to suffer from selenium poisoning, as a result of living on toxic soils or from excess selenium in feed (Giri et al. 2021).

Dietary selenium intake in vertebrates depends on the underlying selenium content and bioavailability of the local environment, where consumed plants grow or animals feed. Aquatic environments generally act as a sink for land selenium and diets of aquatic vertebrates are consequently very high in selenium (May et al. 2008; Sarangi et al. 2017). This results in a vastly different degree of selenium exposure between land and aquatic species. By extension, land and aquatic species encounter drastically different selective pressures surrounding their selenium intake and regulation. Still, soils across the globe can vary a hundredfold in their selenium content (Sarangi et al. 2017), and terrestrial vertebrates may also encounter substantially different levels of selenium in the diet.

The biological role of selenium is mediated via the amino acid selenocysteine (Sec), which is the key residue of selenoproteins. Selenocysteine is the 21st amino acid, only having been identified in 1974 and lacking a clear mechanism of its production and incorporation into proteins until the 1980s (Stadtman 1974; Chambers et al. 1986). Sec is encoded by an in-frame UGA codon, which usually acts as a stop codon (Chambers et al. 1986). However, the presence of a SElenoCysteine Insertion Sequence (SECIS) element, alongside various cofactors, redirects the translation of the UGA stop codon into Sec (Berry et al. 1992). The SECIS structure can be identified in the 3'UTR of the mRNA in selenoproteins in eukaryotes and archaea (Labunskyy et al. 2014), and is often used to identify the Sec codons that most databases otherwise classify as the end of an open reading frame (Romagné et al. 2014; Sarangi et al. 2017).

Selenocysteine mediates the catalysis of selenoproteins, governed by the unique enzymatic properties of selenium. Most of the functionally characterised selenoproteins have roles in redox regulation, whilst the function of many others remain either unknown or not fully elucidated (Mariotti et al. 2012). When knocked out in mice, the loss of selenoproteins can result in death (*e.g.,* in the case of *TR1, TR3* and *GPX4*) or reduced fitness (Matsui et al. 1996; Yant et al. 2003; Conrad et al. 2004; Jakupoglu et al. 2005; Peters et al. 2006; Fomenko et al. 2009), strongly supporting their biological necessity. Selenoproteins have also been suggested to play a role in maintaining immune response, male reproduction and cancer prevention in humans (Arnér and Holmgren 2006; Hatfield et al. 2006; Papp et al. 2007).

Selenocysteine itself is utilised in a range of catalytic redox reactions, including repairing oxidised methionines in proteins, removal of hydroperoxides, regulating activation of thyroid hormones and regulating reductions of thioredoxin (Santesmasses et al. 2020). Often the catalytic ability of selenocysteine is compared to cysteine, its analogous amino acid which differs only its replacement of selenium by sulfur, and is a point mutation away from the Sec codon (Cys encoded by UGC and UGT codons (Sarangi et al. 2017)). Directly substituting Sec for Cys has been shown to reduce the catalytic ability of an enzyme by 5% (Stadtman 1996), reflecting the decreased reactivity and nucleophilicity of Cys (Arnér 2010). The greater catalytic potential of Sec in comparison to Cys has also been suggested to be a result of its increased resistance to oxidation stress (Snider et al. 2013) and its activity across a wider range of pH conditions and substrates (Gromer et al. 2003). Indeed, the unique role of Sec, and low exchangeability between Sec and Cys, is supported by the inferred strong evolutionary constraint acting on selenocysteine in selenoproteins (Castellano et al. 2009).

Still, the exchange of Sec to Cys has been inferred to have occurred numerous times during vertebrate evolution (see **Figure 1.12**), begging the question as to what evolutionary mechanisms allow the loss of such a catalytically powerful residue. Compensatory mutations have been shown to restore catalytic ability, although only at 50% of the catalytic rate of the corresponding selenoenzyme (as for the Thioredoxin reducatase of *Drosophila melanogaster*; (Gromer et al. 2003)). The sulfur-containing Cys may also compensate for its lower catalytic activity via its higher expression than Sec (since the Sec translation is comparatively inefficient; (Liu et al. 2012)) or its escape from the limitations of relying on rare selenium in place of common environment sulfur.

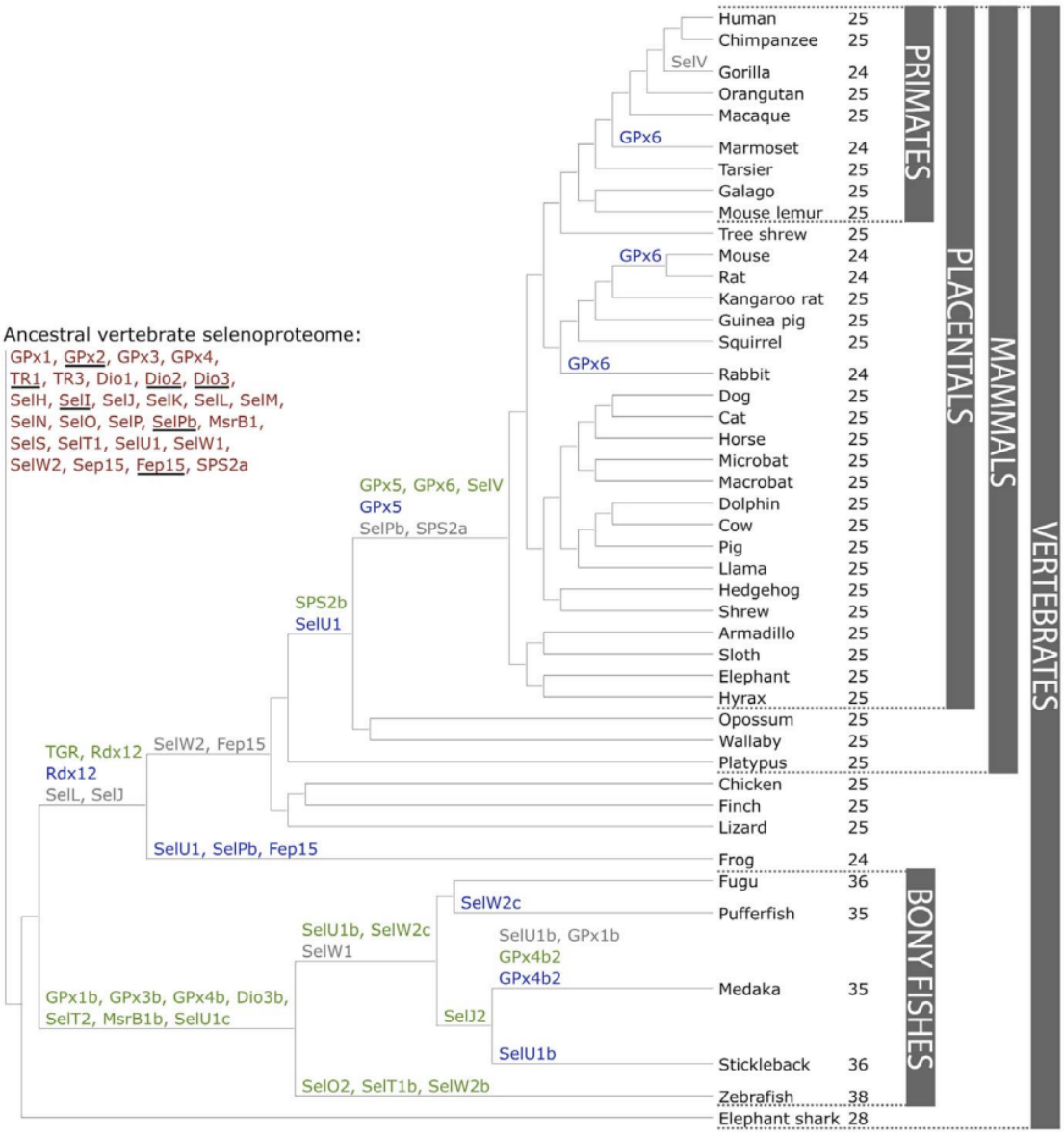


Figure 1.12. Evolution of the vertebrate selenoproteome. *The ancestral vertebrate selenoproteome given in dark red at the root of the tree. Unique selenoproteins in vertebrates are underlined. Novel selenoproteins created by duplications are given in green, loss is given in grey. Exchanges from Sec to Cys given in blue (bar SelW2c in pufferfish, where Sec is replaced by arginine). Number of selenoproteins predicted in each species given on the right. Taken from (Mariotti et al. 2012).*

1.8.1.1. Selenoproteome Diversity

Vertebrate species have selenoproteomes containing 24 to 38 selenoproteins, derived from the a common ancestral selenoproteome of size 28 (Castellano et al. 2009; Mariotti et al. 2012). Comparative analyses of nucleotide and protein sequences have inferred a complex history of exchanges from Sec to Cys in the catalytic site of selenoproteins, as well as multiple selenoprotein duplications, throughout vertebrate history (Mariotti et al. 2012). This includes proteins that have exchanged Sec for Cys in multiple vertebrate lineages (as is the case of *GPX6*, losing Sec many times across mammalian lineages); proteins that were generated through duplications of selenoproteins but now lack Sec in all organisms (*e.g., GPX5, RdX12*; likely losing their Sec residue before the duplicated gene haplotype became fixed); and repeated duplications of selenoproteins in multiple selenoprotein families, with change or gain of function (*e.g.,* in the GPX and TR families) (**Fig. 1.12;** (Mariotti et al. 2012)).

The repeated selenoprotein duplications across bony fish lineages is also highlighted, particularly in the zebrafish, and is inferred to be a result of fourteen distinct events (Mariotti et al. 2012). It has been suggested that the larger selenoproteome in fish is associated to the increased amount of selenium in their aquatic environments (Sarangi et al. 2017). As a consequence of this environmental abundance of selenium, fish may have evolved a greater dependence on selenium, supported by their maintenance of selenium transporting mechanisms in the body (Lobanov et al. 2007; Sarangi et al. 2018).

In summary, selenoprotein diversity has been linked to the broad environmental levels of selenium experienced by divergent vertebrate taxa (particularly contrasting the selenoproteome between terrestrial and aquatic vertebrate taxa (Sarangi et al. 2018)), as well as to the unique catalytic role of Sec (Castellano et al. 2009). Through analysis of selenoproteome size and conservation of individual selenoproteins, the exact selective pressures governing macroevolution of selenoproteins can be explored.

Understanding the role of environmental selenium versus the catalytic role of Sec in shaping genomic diversity across taxa should also integrate selenium-associated evolution at the micro-scale, that is within individual species (as previously discussed within the frame of modern human populations, see **Section 1.7.3**). Whilst it is clear that dietary selenium has uniquely shaped vertebrate evolution, the exact evolutionary dynamics, including those selective drivers, remain an exciting part of evolutionary biology, molecular and population genetics.

Chapter 2: The Power and Limitations of Identifying Local Adaptation in Modern Humans

2.1. Overview

Local adaptation has occurred throughout the evolutionary history of modern humans as a result of the highly varied environments and selective pressures of which our species encounters (Fan et al. 2016; Rees et al. 2020). However, the genomic nature of local adaptation is highly variable in regards to the strength of selection, the origin and number of alleles under selection, and the timing of the onset of selection (see **Chapter 1**). Hence, local adaptation cannot be solely characterised by strong, uniform signatures of positive selection (historically often described as a "hard sweep"). Instead, local adaptation is likely also accompanied by weaker signatures of positive selection: those left by selection on standing variation or by selection on multiple genes, as likely in complex trait adaptation (historically often described as a "soft sweep"; (Pritchard et al. 2010; Peter et al. 2012; Hermisson and Pennings 2017)). Many of the current methods to identify the signatures that positive selection leaves on the genome (see **Chapter 1**) are poorly-equipped to identify these subtler signatures, and it is unclear which methods are the most powerful in identifying local adaptation mediated by selection on, for example, standing variation or multiple genes.

Here I explore the power of different approaches to identify the genomic signatures of soft sweeps, including new methods that have not extensively been tested. To do so, I design a simulation framework that models local adaptation on one of four major human populations using SLiM (Haller and Messer 2019), modelling weak selection occurring on segregating alleles at one of four timepoints (1kya, 5kya, 10kya and 40kya). I then test the accuracy of allele-frequency differentiation, haplotype-based and tree-recording methods to identify such instances of positive selection. I also test their power to identify polygenic selection by comparing these methods in the gene set method $SUMSTAT$ (Daub et al. 2013). I show the high power of the allele-differentiation statistic F_{ST} and tree-recording method *Relate* in identifying local adaptation as recent as 10,000 years old, both at the monogenic and polygenic level. On the contrary, I show that the power of haplotype-based statistics is insufficient in identifying selection events mediated by weak selection on standing variation.

2.2. Background

Positive natural selection drives adaptive evolution across all organisms, increasing the frequency of traits that convey a fitness advantage. These traits may be fixed across a species or vary over populations and individuals, often correlating with local selective pressures (Darwin and Wallace 1858; Savolainen et al. 2013). Human adaptation to such local environmental pressures, hereby referred to as local adaptation, has been shown to play a role in the modest genetic and phenotypic differentiation that exists between populations (Key et al. 2018; Rees et al. 2020). Most notably this includes adaptations in response to local diet, hypoxia, temperature, UV levels and pathogen load, amongst others (see **Chapter 1;** (Lamason et al. 2005; Norton et al. 2007; Tishkoff et al. 2007a; Genovese et al. 2010; Yi et al. 2010; Jacobs et al. 2013; Bigham and Lee 2014; Vernot and Akey 2014;

Fumagalli et al. 2015; Schlebusch et al. 2015; White et al. 2015; Minster et al. 2016; McManus et al. 2017; Key et al. 2018)).

Often, human populations are exposed to novel selective pressures when migrating into new environments (particularly when those environments are extreme, such as those at high altitude or with extreme temperatures (Ilardo and Nielsen 2018)). Whilst modern humans have long since inhabited variable African environments (inhabited since the origin of our modern species approximately 200,000 years ago, (White et al. 2003; Dusseldor et al. 2013)), many global environments were only colonised following the "Out of Africa" migration (50-70,000 years ago, (Soares et al. 2012; Haber et al. 2019)). This preceded major human expansions to Oceania (Bowler et al. 2003), Eurasia (Fu et al. 2014; Seguin-Orlando et al. 2014) and the Americas (Raghavan, DeGiorgio, et al. 2014; Raghavan, Skoglund, et al. 2014; Rasmussen, Anzick, et al. 2014). In even more recent time, the emergence of novel cultural practices has also resulted in the rapid exposure of novel selective pressures, such as those emerging following the agricultural revolution approximately 10,000 years ago.

2.2.1. Genomic Signatures of Local Adaptation

The timepoint of selection is one factor that contributes to the genomic signatures of positive selection. These signatures also largely depend on the mode of adaptation, such as the origin of the selected allele or the degree of polygenicity, where some modes result in subtler signatures that are more challenging to identify (see **Chapter 1**). When local adaptation occurs in populations with extreme demographic histories, such as bottlenecks or partially resolved admixture events, these signatures are more elusive still as they can be masked by neutral processes that may appear as under selection, or we may simply lack an understanding of how signatures may present under such histories (Peter et al. 2012; Gopalan et al. 2022).

To understand the signatures of positive selection, selection has historically been categorised as either a "hard sweep" or "soft sweep" (Pritchard et al. 2010; Peter et al. 2012; Hermisson and Pennings 2017). The "hard sweep" model describes strong selection on a *de novo* mutation which results in the rapid increase of frequency of the advantageous allele, together with a battery of signatures of positive selection (Pritchard et al. 2010; Schrider and Kern 2016). Whilst this long underpinned classic ideas of selection, many have suggested the importance and prevalence of "soft sweeps" in human evolution (Hermisson and Pennings 2005, 2017; Prezeworski et al. 2005; Pritchard et al. 2010), of which recent studies have demonstrated (Schrider and Kern 2016, 2017). These "soft sweeps" are the result of slow increases in allele frequency due to weak selection, or selection that acts on already segregating alleles (selection on standing variation, or SSV (Hermisson and Pennings 2005, 2017; Peter et al. 2012)).

SSV has been proposed to be a particularly likely mode of selection in local adaptation, especially if the allele has been maintained in the population due to balancing selection (and therefore necessarily affects phenotype and fitness (Andrés 2011; Rees et al. 2020)). However, the signatures of SSV can be particularly difficult to identify. Since the mutation is evolving under drift before the onset of selection, which may indeed be the major proportion of the mutation's lifetime, the adaptive allele is likely present on diverse genetic backgrounds and therefore lacks the signatures of linked variation that accompany "hard sweeps" (see **Chapter 1**).

It is also expected that polygenic adaptation may be common in human local adaptation (indeed, with evidence to suggest so (Hancock, Alkorta-Aranburu, et al. 2010; Daub et al. 2013, 2013; Berg and Coop 2014; White et al. 2015)) since many complex traits are polygenic in nature. Polygenic selection is driven by small shifts in allele frequency which occur across groups of phenotypically-related genes (those that all contribute to the same phenotype) and interact to shift the phenotype in the adaptive direction (Le Corre and Kremer 2012). Polygenic adaptation thereby leaves many, weak signatures along the genome, and is challenging to uniformly characterise. The degree of polygenicity varies across traits, with genes responsible for a phenotype potentially ranging in number from few to thousands (Daub et al. 2013; Berg and Coop 2014; White et al. 2015; Zhang et al. 2015; Boyle et al. 2017; Mathieson 2021). Further, genes associated with a trait may not all respond similarly to selection due to differences in effect size (Berg and Coop 2014; Mathieson 2021), and some genes may show stronger, almost monogenic signatures of positive selection (Wagner and Zhang 2011; Fraïsse et al. 2019). This may also be the case under traits where many of the functionally associated genes have deleterious pleiotropy, resulting in selection acting on few alleles (Chevin and Hospital 2008).

2.2.2. Identifying Signatures of Local Adaptation

Hence, identifying local adaptation can often become a quest to identify subtle and variable signatures of positive selection across an unknown number of genes. Many methods identify particular aspects of the signatures of positive selection, summarised into a single statistic. These can then be used to identify the loci which show outlier values according to the empirical background of the genome, and hence are the most likely candidates for selection. Commonly used statistics summarise allele frequency differentiation (Weir and Cockerham 1984), haplotype length (Voight et al. 2006; Sabeti et al. 2007; Ferrer-Admetlla et al. 2014; Szpiech et al. 2021) or patterns of the site frequency spectrum (Tajima 1989; Excoffier et al. 2013). However, many of these classical methods have been designed to identify strong, monogenic signatures of positive selection (Sabeti et al. 2006; Pritchard et al. 2010; Hermisson and Pennings 2017), and may lack the power in identifying the signatures that accompany SSV. Methods based in allele frequency differentiation are the broad exception to this and can be used with much success in identifying SSV, since they do not rely on linked variation (Weir and Cockerham 1984; Yi et al. 2010; Yassin et al. 2016; Crawford et al. 2017; Librado and Orlando 2018; Schmidt et al. 2019).

Tree-recording methods (Rasmussen, Hubisz, et al. 2014; Kelleher et al. 2019; Speidel et al. 2019; Hubisz and Siepel 2020) show increasing promise to identify the subtler signatures of selection that accompany both SSV and polygenic selection. These methods can be used to build individual genealogies along the whole length of the genome, and reconstruct the evolutionary history of each site (where histories differ according to recombination break points). In theory, positive selection can then be identified not by way of summarising the genealogy or evolutionary history into a statistic, but by direct inference from the genealogy itself. By avoiding the collapse of complicated evolutionary patterns into a single statistic, albeit remaining an inference, it is likely that these methods are better suited to identifying weaker signatures of positive selection that are not characterised by rapid allele frequency change or unusually long haplotypes. Indeed, tree-recording methods have already been shown to have success in identifying both monogenic and polygenic selection in humans (Kelleher et al. 2019; Speidel et al. 2019).

Methods to identify polygenic selection often rely on a good understanding of the genetic bases of phenotypic traits, such as the effect sizes of each SNP, as estimated by GWAS, on the candidate trait under selection (Berg and Coop 2014; Field et al. 2016; Berg, Zhang, et al. 2019; Zeng et al. 2021). Methods which integrate such effect sizes (*e.g.,* searching for alleles with similar effects and positive covariance (Berg and Coop 2014; Berg, Zhang, et al. 2019)) are far the most common when identifying polygenic selection, but they have been shown to overestimate the signature of polygenic adaptation if population stratification is not fully accounted for (Berg, Harpak, et al. 2019; Sohail et al. 2019). Gene set methods, those that combine the signatures from multiple genes within a functional set (such as biological pathways; (Subramanian et al. 2005; Daub et al. 2013, 2017)), do not require the same trait information of a population, and are potentially more robust to biases emerging from hidden population sub-structure. Some such methods, *e.g., SUMSTAT* (which simply sums summary statistic across the gene set, see **Section 1.6.5.2;** (Daub et al. 2013)), can integrate any summary statistic. This makes them highly customisable and open to using selection statistics that have been shown to be more powerful under the hypothesised dynamics of selection of which they are being used to investigate.

It is important to understand the power of each of these aforementioned methods to identify positive selection, particularly under modes of selection that leave subtle signatures on the genome of which they were not designed to detect. Hence, it is especially important to ask how power to detect SSV varies between methods, and how it may be affected by the strength of selection, in combination with varied timepoints of selection and population histories. Whilst this is imperative for all methods, it is especially pertinent for recent methods in the field, such as tree-recording methods, since their power and limitations in identifying the genomic signatures of positive selection have not been widely explored. In regards to identifying polygenic selection, it is further interesting to ask how the variance of signatures of positive selection, and the number of trait-associated genes acting under selection, may affect the power of commonly used methods.

2.2.3. Study Overview

Here, I build a simulation framework to assess the power of methods to identify local adaptation. Using the forward simulator SLiM (Haller and Messer 2019), I model selection on a 100kbp genomic region, where selection is both weak and acting on already segregating variants (SSV). I model the demographic history of four major global populations and specify that selection occurs locally on one population at one of four timepoints (1kya, 5kya, 10kya, 40kya; thereby testing the power to detect signatures left by recent selection, selection surrounding agricultural change and selection surrounding major migrations to new environments).

I test the power of the recent tree-recording *Relate* method (Speidel et al. 2019) against five traditional neutrality tests, including those based on allele frequency differentiation (F_{ST} (Weir and Cockerham 1984)) and haplotype-length (iHS, nSL, $XPEHH$ and $XPnSL$ (Voight et al. 2006; Sabeti et al. 2007; Ferrer-Admetlla et al. 2014; Szpiech et al. 2021)). Here, the haplotype-based methods are expected to have low power (given that they were not designed to identify selection on standing variation) and instead serve to contextualise the power of *Relate* and its comparison to F_{ST}. I then integrate these individual methods into the gene set method $SUMSTAT$ (Daub et al. 2013), and test the power of identifying local, polygenic selection. I show high power of two methods to

identify local adaptation in modern humans: F_{ST} and the $Relate$, including when integrated into the gene set method $SUMSTAT$. To my knowledge, this study is the most comprehensive exploration to date of the power of $Relate$ to identify positive selection (as well as of the more recently developed nSL and $XPnSL$ statistics) and allows valuable insight into the power and limitations of identifying local adaptation when using top-performing methods.

2.3. Methods

2.3.1. Simulation Design

2.3.1.1. The Genomic Model

The forward-simulator SLiM (Haller and Messer 2019) was used to simulate genomic segments of approximately 100,000 base pairs. Each segment was initiated with random nucleotides across its length and included exon, intron and non-coding regions (organised according to the SLiM guidance (Haller and Messer 2019)). Variable recombination rates were also specified across this region, modelled according to the inferred distribution of recombination rates in the human genome (as calculated from chr15 of the 929 individuals of the HGDP dataset (Bergström et al. 2020), see **Fig S2.1**), with 100 different recombination rates given across this region. According to this distribution, and in line with the relevant literature (Barroso et al. 2019), a gamma distribution of mean 1.311 and shape parameter 0.509 was used to draw recombination rates. The mutation rate was uniform throughout the 100kb region, specified as 1.25×10^{-8} per generation and following the Jukes-Cantor model.

2.3.1.2. The Demographic Model

I simulate the demographic history of four metapopulations: African, European, East Asian and American. This model is the combination of two pre-existing demographic models, one which represents the history of African, European and East Asian populations (Gravel et al. 2011) and one which exclusively models American demographic history (Gravel et al. 2013). I integrate the inferred demographic history of the Puerto Rican population from the latter into the former model, and use as the proxy of an American population (see **Fig. 2.1**).

These simulated populations broadly approximate the demographic history of each metapopulation but do not accurately represent the demographic history of every individual population within that metapopulation. The model does however include major bottlenecks present in the history of each metapopulation, and therefore approximates the breadth of demographic history amongst modern human populations.

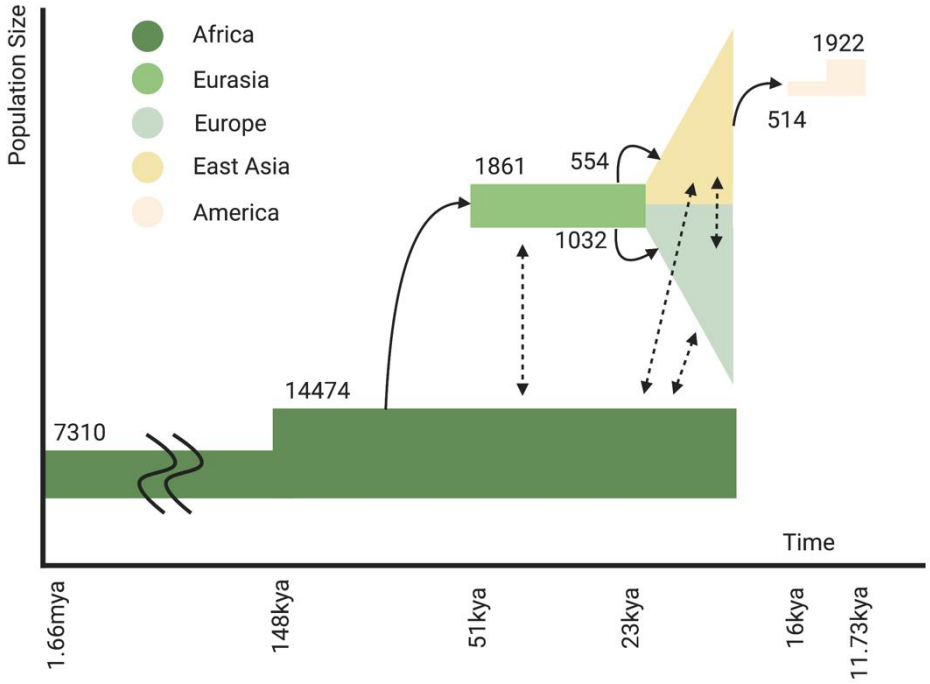


***Figure 2.1**: Schematic illustration of the demographic model used in the simulations.* *A combination of the demographic models from Gravel et al. 2011, 2013. showing the demographic histories of African, European, East Asian and American populations.*

2.3.1.3. Initiation of Selection

The onset of selection was set at one of four timepoints (1kya, 5kya, 10kya and 40kya) in only one of the four metapopulations. A single polymorphic allele segregating in the focal population is tagged and given a selection coefficient drawn from a uniform distribution between 0.001 and 0.005. This tagged allele must be within the middle 10,000 bp of the simulated 100kb genomic region to ensure haplotype information is not lost at the edges of the region. The tagged allele must also be at a frequency between 0.1-0.15 at the onset of selection to decrease the probability that it is lost due to drift. If no suitable allele exists, or the allele is still lost during the simulation's subsequent run, the simulation is terminated and restarted using the next available seed.

This model simulates weak, variable, selection acting on previously existing genetic variation (SSV). Each successful simulation run can then be used as a proxy for weak selection acting on a single genomic region, analogous to a gene region or haplotype. Polygenic selection can also be modelled by grouping multiple simulation together in sets (as a set of "loci"), where selection coefficients are weak and variable across loci.

2.3.2. The Simulation Run

An initial burn-in period was simulated between 1.66mya and 70kya to allow the ancestral population to reach mutation-drift equilibrium. Here, beneficial, neutral or deleterious mutations were initiated in the exon regions, where the selection coefficients

of deleterious mutations were drawn from a gamma distribution (mean: -0.03 and shape parameter: 0.2; (Boyko et al. 2008; Kim et al. 2017)) and beneficial mutations drawn from an exponential distribution (mean: 0.01, capped at 0.05; (Orr 2003; Brajesh et al. 2019)). Neutral mutations also can appear in the intron and non-coding regions.

To reduce CPU time, simulations were rescaled by a factor 5 to reduce the number of simulated individuals (reducing 7310 individuals to 1462). Here, μ, r and s (mutation rate, recombination rate and selection coefficients respectively) were scaled up whilst N_e was scaled down, which maintains the necessary population-genetic parameters of $N_e\mu, N_e r$ and $N_e s$ ((Liu et al. 2010; Lynch and Ho 2020) and expected site frequency spectrum; see **Fig S2.2**). The generational time was also down-scaled by the same factor to account for the fact that genetic drift occurs faster in smaller populations (Kimura and Ohta 1969). This rescaling reduced the CPU time by over a factor 20.

The VCF output from the burn-in was then expanded to 14,474 individuals via random mating in the first generation. This represents the ancestral African population at 70kya, which then undergoes population splits, expansions and migrations as described in **Fig. 2.1**. During this stage of the simulation, only a singular beneficial mutation is initiated at one time-point in one of the four metapopulations (the focal population). This eliminates the risk of stochastically occurring positive selection events on untagged mutations masking the focal selection events.

For each scenario (the combination of one selection timepoint in one metapopulation), 10,000 simulations were run on the requirement that the tagged mutation remains polymorphic in the focal population. For each run, VCF files of 50 individuals for each metapopulation were generated as output, alongside the position, selection coefficient and final frequency of the tagged mutation. A CSV files was also generated containing information on the inclusive upper bound position of each recombination rate, which was converted to a standard genetic map format, where recombination rate is given in cM/Mb, using the following formula:

$$gpos_{n+1} = \frac{(ppos_{n+1} - ppos_n \times rrate_{n+1})}{10^6} + gpos_n$$

2.3.3. Use of Simulation Output

2.3.3.1. Application of Methods to Identify Selection

I chose to apply six methods to identify the genetic signatures of positive selection. Four of these methods use haplotype structure to infer SNPs with evidence of positive selection, but do so in subtly different ways. iHS and nSL both consider the length of haplotype homozygosity (where extended haplotype homozygosity is indicative of alleles rapidly rising in frequency, as expected under strong positive selection), but iHS measures length as the recombination distance, whereas nSL measures length as the number of segregating sites (Voight et al. 2006; Ferrer-Admetlla et al. 2014). nSL is an edit of iHS, a commonly used method to identify positive selection, and has been suggested to be more powerful when detecting selective sweeps on standing variation (Ferrer-Admetlla et al. 2014). $XPEHH$ and $XPnSL$ are extensions of these two methods (of iHS and nSL respectively), and compare the haplotype homozygosity between two populations to identify SNPs with unusually long haplotype length in one population (Sabeti et al. 2007;

Szpiech et al. 2021). The remaining two methods used here include one which uses allele differentiation between populations to identify evidence of positive selection (F_{ST}; (Weir and Cockerham 1984)) and one that uses the inferred trajectory of an allele through its history to infer the probability of positive selection (*Relate* (Speidel et al. 2019)). *Relate* first infers local trees along the genome (where unique trees are separated by recombination breakpoints); it uses a Hidden Markov Model to reconstruct a chromosome as a mosaic of other samples and iteratively clusters the samples most likely to have been copied from each other together (resulting in the final inferred tree; (Speidel et al. 2019). It then maps mutations onto each tree and simultaneously estimates branch lengths, mutation rates and effective population size to re-infer the trees, which can then be used to estimate effective population sizes of subpopulations, cross-coalescence rates between populations and the likelihood of a variant's trajectory under neutrality.

All haplotype-based statistics were calculated using the *SELSCAN* programme (Szpiech and Hernandez 2014) and normalised according to SNP frequency. For *XPEHH* and *XPnSL*, calculations were repeated for each combination of focal population with the three remaining populations. *VCFTOOLS* (Danecek et al. 2011) was used to calculate F_{ST} according to the Weir and Cockerham (1984) method (Weir and Cockerham 1984), again repeated for combinations of focal population with the three remaining populations. The *Relate* programme (Speidel et al. 2019) was ran according to suggested default parameters and used to calculate the probability of a variant's trajectory (analogous to a p-value to indicate selection), given its inferred genealogy.

For the calculation of haplotype-based statistics, the "*--trunk-ok*" parameter was used in *SELSCAN* (Szpiech and Hernandez 2014), which specifies that the statistic should still be calculated despite the extended haplotype homozygosity failing to decay to the suggested threshold of 0.05. This is due to the discrete size of the simulated genomic region, and results in the data in the end tails of haplotype decay being lost. I compared the distribution of *iHS* data from the initial simulated 100kbp genomic regions under selection to 300kbp simulated genomic regions under selection and of matched seed, and observe the distributions not to be statistically different (Wilcoxon test; $Z = 1.361414, pvalue = 0.1734$), concluding that the length of the simulated genomic regions does not significantly affect the calculation of the haplotype-based statistics.

Finally, I use the gene-set enrichment method *SUMSTAT* (Daub et al. 2013) as the method to identify polygenic selection. Gene sets of various sizes (10, 20, 40, 60) were built by random sampling of simulated gene regions, and varying the proportion of gene regions under selection compared to neutrality (20%, 40%, 60%, 80% and 100% gene regions under a selection), hereafter referred to as polygenic adaptation gene sets. Following the calculation of the test statistics above, the strongest score (in the direction of selection) for each gene region is taken and summed across gene sets. Hence, this method considers the signatures of positive selection on potentially small effect mutations across the entire gene set, and has been shown to be more powerful than gene set enrichment analysis in identifying polygenic selection (Tintle et al. 2009).

2.3.3.2. Isolating Signatures of Positive Selection

I use empirical neutral distributions, built from the output values calculated on neutral simulations, to identify SNPs with evidence of positive selection for each of the six methods. To build these distributions, I use the same burn-in simulations and consequent simulations (using the seed numbers of each successful simulation run, see **Section**

2.3.2), but with no onset of positive selection in any metapopulation. I then apply the same six methods on these neutral simulations, and build a distribution from the subsequent output values (normalised where appropriate, see **Section 2.3.3.1**) for 10,000 of these neutral simulations. SNPs are identified as having evidence of positive selection, according to each method separately, if they fall in the 5% tail of the empirical neutral distribution for the respective method. The potential for bias in this methodology is recognised, given that I do not condition on the maintenance of the focal SNP being at same frequency in the neutral simulations as under the simulations including selection.

I use an analogous method to build the neutral distribution corresponding to $SUMSTAT$ values; I generate 1000 random gene sets (for each gene set size of 10, 20, 40 and 60), calculate the $SUMSTAT$ value across these gene sets as described in **Section 2.3.3.1**, and from these values build the neutral distribution. Gene sets with $SUMSTAT$ values in the 5% tail of these empirical neutral distributions are similarly assigned evidence of positive selection, again separately for each method integrated into the $SUMSTAT$ framework.

I also evaluate the use of a neutral distribution to identify SNPs using the $Relate$ method (Speidel et al. 2019). This programme outputs the probability of positive selection in the form of a $-log10pvalue$ (where the $pvalue$ corresponds to the probability of a variant spreading to its modern observed frequency) and previous work has explicitly used this transformed $pvalue$ as evidence of selection (Speidel et al. 2019). Here, I ask if using the tails of the empirical neutral distribution to identify SNPs with extreme $-log10pvalues$ as those with evidence of selection is more accurate, as well as if it reduces the difference in power between populations of differing demographic histories (see **Section 2.4.1**).

2.3.3.3. Evaluating Accuracy of Methods to Identify Positive Selection

I evaluated the accuracy of each method to identify monogenic selection in three ways:

1) By calculating the percentage of true selected SNPs that fall in the 5% empirical tail of the neutral distribution (as described in **Section 2.3.3.1**).
2) By calculating the percentage of SNPs with the strongest evidence of selection (most extreme statistic value) within a simulated gene region that is the true selected SNP;
3) By calculating the average distance from the SNP with the strongest evidence of selection to the selected SNP (where physical distance can be treated as an approximation of genetic distance, but should not be considered synonymous).

Therefore, I asked the suitability of each method to:

1) Identify SNPs under selection as showing evidence of selection (according to the 5% empirical tail)
2) Identify SNPs under selection as the most likely candidate of selection
3) Identify the area of the gene region under selection

The $SUMSTAT$ method (Daub et al. 2013) was evaluated by assessing the percentage of polygenic adaptation gene-sets that fall in the 5% tail of the empirical neutral distribution, as described in **Section 2.3.3.1.** This was repeated for all conditions (all gene-set sizes and proportions of genes under selection in the gene set), and allows the evaluation of which proportion of a gene set under selection results in appreciable power.

I caution that whilst the accuracies calculated for the European, East Asian and American populations are useful for observing patterns across methods and time, these populations

should not be directly compared to each other. This is because I condition on the selected allele to persist to the end of the simulation. While this is a necessary condition widely used in comparable power analyses, it does result in differential biases across populations (since differences in demography result in differences in the probability of an allele to survive to the time of sampling). Therefore, the final set of simulated genomic regions is thus informative but not perfectly comparable across demographic histories and populations.

2.4. Results

2.4.1. Optimising the Relate Method

The *Relate* programme (Speidel et al. 2019) outputs the probability of a variant spreading to its modern observed frequency in the form of a $-log10pvalue$, where a value lower than -1.30103 indicates a probability, or p-value, of less than 5% and can be taken as evidence of selection. However, power is not independent of the effective population size of the population, and hinders comparisons across populations. I therefore test if using the tails of a neutral distribution to identify selected SNPs 1) results in an increase in power in some populations and 2) decreases the differences in power across different populations. To do so, I compared the number of selected SNPs that were below the $-log10pvalue$ threshold of -1.30103 to those identified using the 5% tail of the empirical neutral distribution (built from the $-log10pvalues$ calculated from neutral simulations; see **Section 2.3.3.2**) for selection initiated in each metapopulation at each timepoint.

On average, there is higher accuracy (defined here as the number of selected SNPs identified, akin to a true positive rate or sensitivity) when identifying the selected SNPs using the tail of the empirical neutral distribution, rather than using the raw computed p-values of *Relate* (**Fig. 2.4.1**). This does not remove the difference in power across populations, the highest power is still observed when identifying selected SNPs in the African metapopulation, but does decrease the power differences across populations. Hence, the neutral distribution of *Relate* should be used to identify candidate SNPs and I opt to apply this approach when evaluating the power of *Relate* in downstream analysis.

Power is also not independent of the sample size of the population, with lower sample sizes reducing the power of methods to identify selection (Subramanian 2016; Serdar et al. 2021). Hence, I also suggest that when using smaller sample sizes, using the tails of an empirical neutral distribution to identify selected SNPs may more notably increase the accuracy in comparison to using the raw computed p-values. To assess this, I carried out an additional analysis; I compared the accuracy of both methods to identify selected SNPs when decreasing the sample size from 50 individuals (as used in all following analysis) to 25 individuals. I observe, as expected, that the relative increase in accuracy when using the empirical neutral distribution to identify selected SNPs is higher with the smaller sample sizes. For example, for selection acting at 40kya in the African population, using the neutral distribution increases the accuracy by 16.9% for a sample size of 50 individuals and by 24.8% for a sample size of 25 individuals (**Fig. 2.2**). This further demonstrates the higher accuracy of *Relate* when using the tail of the neutral distribution to identify SNPs under selection and this approach would be suitable when using smaller sample sizes.

However, here I have only evaluated the number of selected SNPs within the tail of the empirical neutral distribution, and the tail will also falsely identify neutral sites as those

under selection. Still, these tails are enriched in true targets of selection, and correctly identify just below 50% of selected SNPs (using the 5% tail) in the best scenario (**Fig 2.2**).

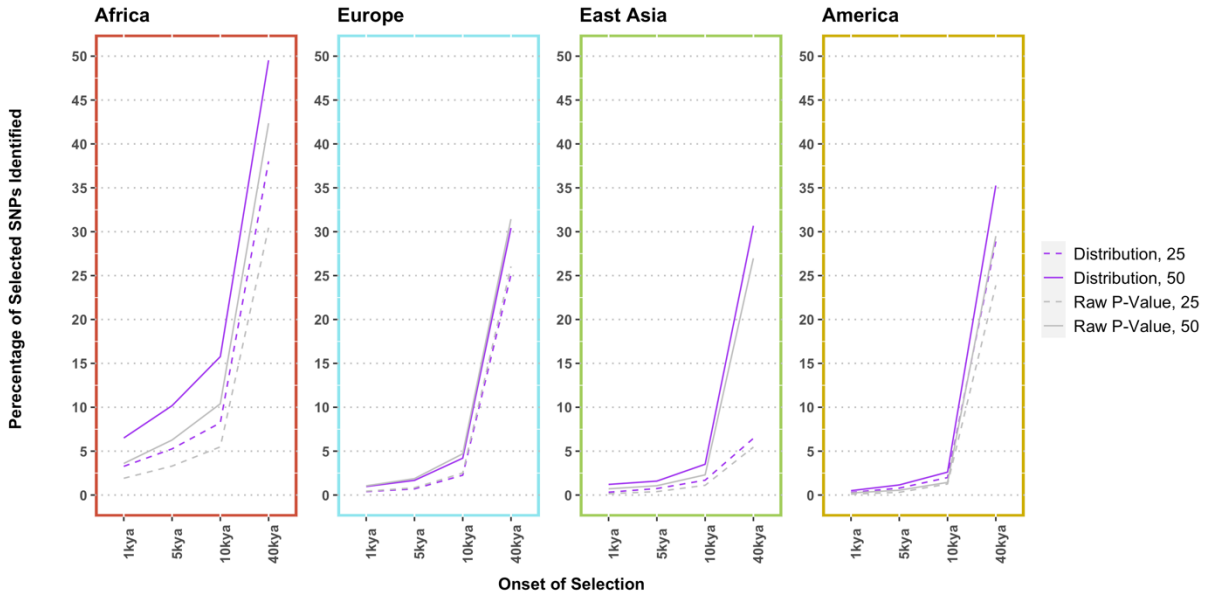


***Figure 2.2: Selected SNPs identified as under selection according to two methods**. The percentage of selected SNPs that are identified as under selection (acting at timepoints 1kya, 5kya, 10kya and 40kya) as defined by falling in the 5% tail of the neutral distribution (Distribution) and as defined by the raw computed p-values of Relate (Raw P-Value), given for samples sizes of 25 and 50 individuals.*

2.4.2. Identifying Monogenic Selection

I now evaluate the accuracy of each method to identify monogenic selection in three ways, as outlined in **Section 2.3.3.3**. I first evaluated the ability of each method to identify selected SNPs as those with evidence of selection (*i.e.,* lying in the 5% tail of the empirical neutral distribution for each method, enriched in true targets of selection). I hence calculated the percentage of selected SNPs within this tail for each method, including *Relate* as informed from **Section 2.4.1**.

The highest accuracy, strikingly so, was obtained when using the *Relate* and F_{ST} methods to identify positive selection (**Fig. 2.3, Fig. 2.4**). The haplotype-based methods show highly reduced accuracy in comparison, but it appears that the accuracy of haplotype-based methods which use the number of segregating sites as a proxy for distance (nSL and $XPnSL$) is higher than those based on recombination distance (iHS and $XPEHH$). By measuring the haplotype length in terms of segregating sites rather than recombination distance, nSL and $XPnSL$ are more robust to recombination rate variation (with iHS shown to be biased towards identifying outliers in regions of low recombination (Voight et al. 2006)). Moreover, using the number of segregating sites incorporates more information on the local genealogy (Ferrer-Admetlla et al. 2014), and

these methods are also somewhat more robust to varying demographic histories of populations.

I also highlight two main observations true for all methods: the highest accuracy is for the oldest selection simulated (selection initiated at 40kya) and for selection identified in African individuals (shown in **Fig. 2.3, Fig. 2.4**). Indeed, this is as expected; both recent selection and selection acting in populations with reduced N_e is typically harder to identify (Field et al. 2016; Subramanian 2016; Serdar et al. 2021). Further, for the cross-population statistics ($F_{ST}, XPEHH, XPnSL$; see **Fig. 2.4**), the accuracy is higher when comparing populations with more recent population splits, most likely reflecting the reduced noise from neutral genetic differentiation in the empirical neutral background (da Silva Ribeiro et al. 2022). Finally, the overall low percentage of tagged variants identified as under selection is noted. Given the small selection coefficients modelled (as low as 0.001 and only as high as 0.005), the extremely recent selection modelled in some cases (1kya or 5kya) and the unsuitability of some methods in identifying selection on standing variation (*i.e.,* haplotype-based methods), this is reasoned as somewhat expected. These simulations are most useful in specifically comparing the *Relate* method to the well-established F_{ST} statistic, where the haplotype-based methods provide an expected lower limit of power for selection on standing variation.

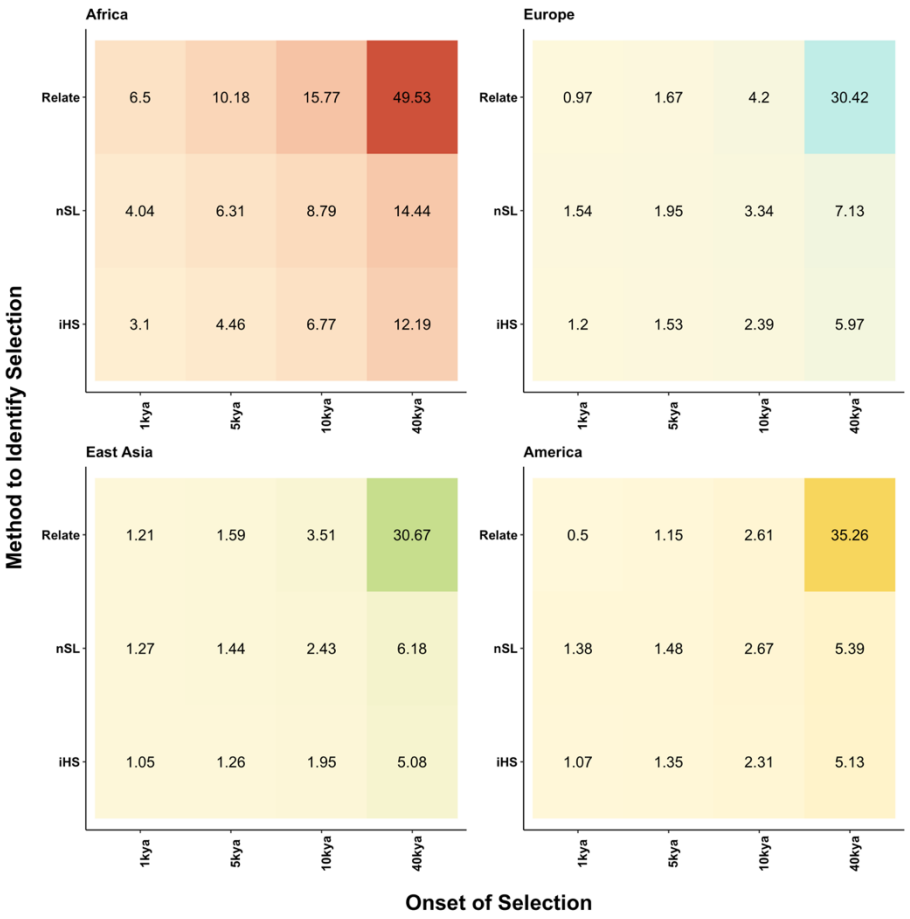


***Fig. 2.3**: **Tagged variants identified as under selection.** The percentage of tagged variants that are identified as under selection (acting at timepoints 1kya, 5kya, 10kya and 40kya in the African (red), European (blue), East Asian (green) or American (gold)*

population), as defined by falling in the 5% tail of the neutral distribution of iHS, nSL and $Relate$.

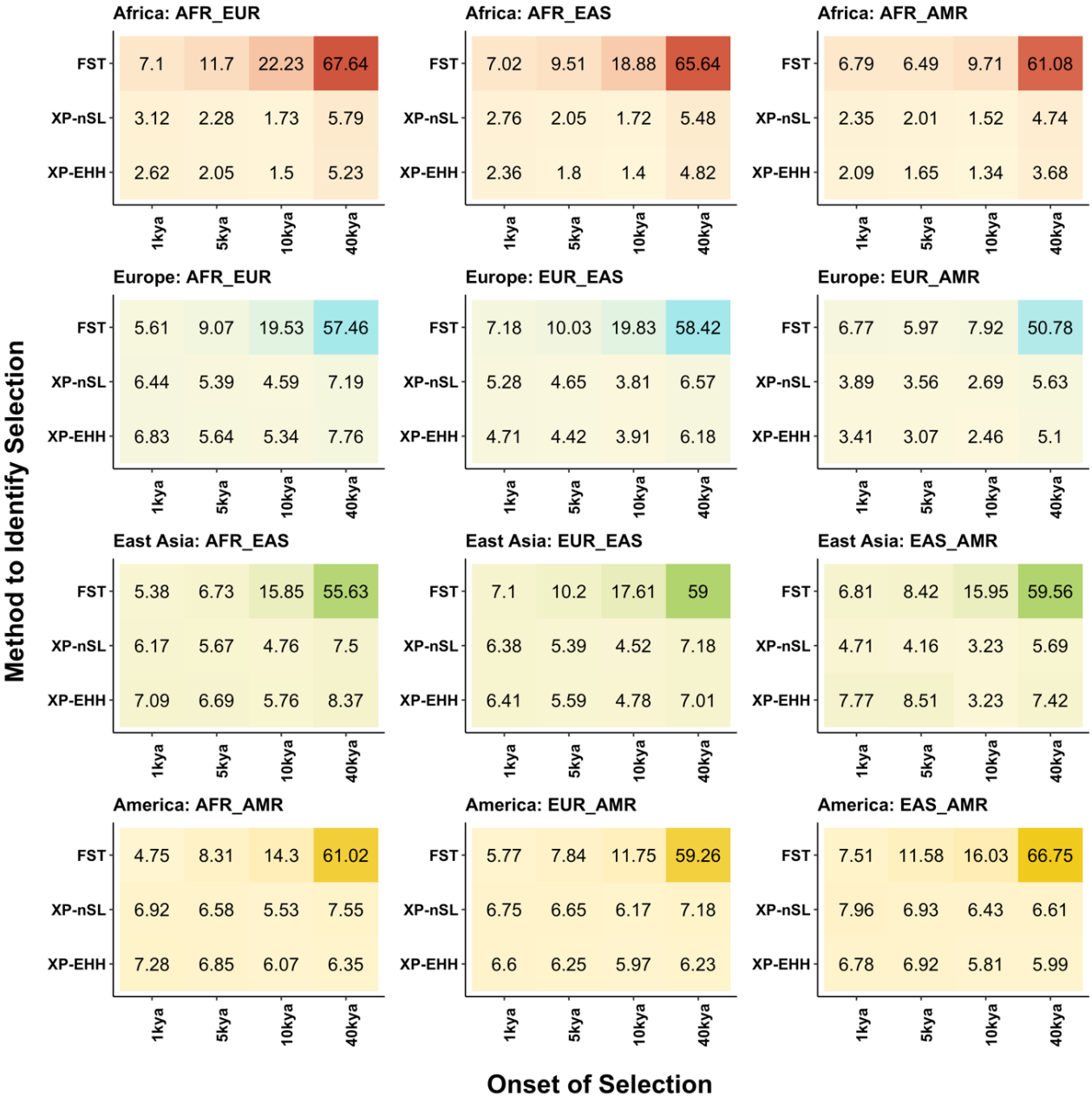


***Fig. 2.4**: **Tagged variants identified as under selection for cross-population methods.** The percentage of tagged variants that are identified as under selection (acting at timepoints 1kya, 5kya, 10kya and 40kya in the African (red), European (blue), East Asian (green) or American (gold) population), as defined by falling in the 5% tail of the neutral distribution of the cross-population statistics $XPEHH, XPnSL$ and F_{ST} (given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America).*

Given that the tails of empirical neutral distributions contain both selected and neutral targets (despite a general enrichment of selected targets), the SNPs with the strongest evidence of positive selection may instead be isolated as the strongest candidate SNPs. Hence, I now evaluate the ability of each method to identify the selected SNP as that with the strongest evidence of positive selection. To do so, I calculate the percentage of selected SNPs with the strongest evidence of selection, that with the most extreme outlier value of the calculated statistic, within each simulated gene region.

The highest accuracy to identify positive selection is also observed here when using the *Relate* and F_{ST} methods, with the percentage of selected SNPs identified again highest for selection acting on the African metapopulation (shown in **Fig. 2.5**). For selection acting on all metapopulations, this accuracy is also again highest when selection acts further back in time (corresponding figures for selection on European, East Asian and American populations now shown in as supplementary figures since they display the same patterns as shown for the analysis of selection on the African population; **Figs. S2.3-5**).

However, the highest percentage of selected SNPs identified as showing the strongest evidence of positive selection according to all methods is only at 10.69% (according to *Relate* when selection is acting at 40kya and in the African metapopulation). This demonstrates the difficulty and general inaccuracy of any of these methods to identify SSV at the exact site, using this approach.

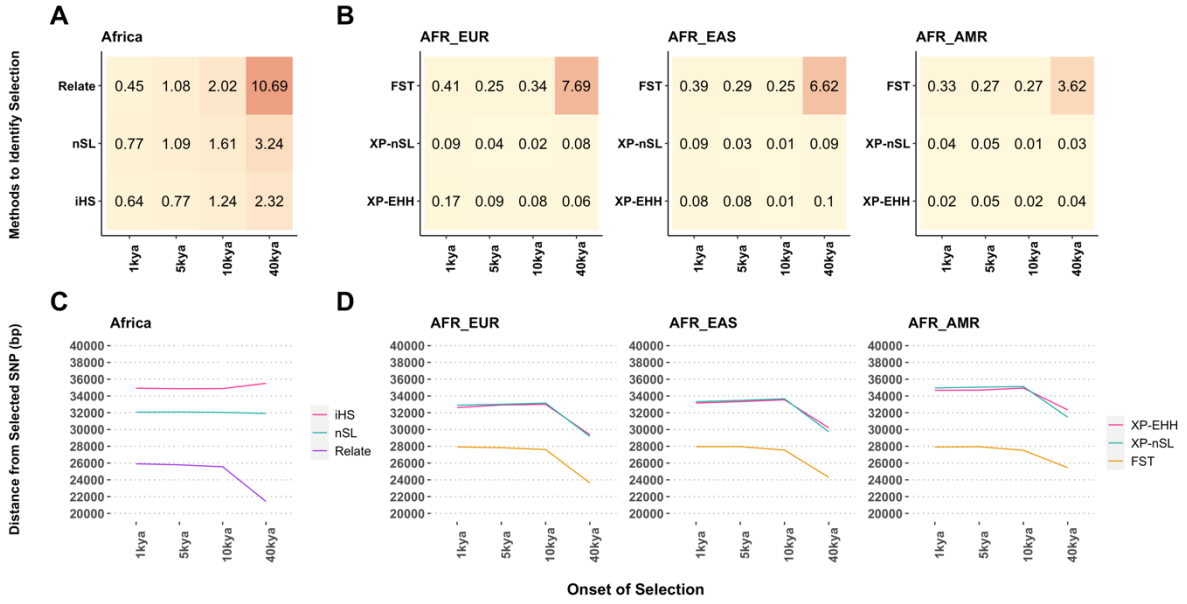


***Fig. 2.5**: **Further analysis of methods identifying selection**. Top panel shows the percentage of tagged variants that are the SNP with the strongest evidence of positive selection across timepoints in the African population for A) iHS, nSL* and *Relate* and B) the *cross-population statistics XPEHH, XPnSL and F_{ST} (given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America). Bottom panel shows the average distance between the tagged variant and the top-ranking SNP for C) iHS, nSL* and *Relate* and D) the *cross-population statistics XPEHH, XPnSL and F_{ST}.*

Finally, and as an extension from the analysis preceding, I then evaluate the accuracy of these methods to identify positive selection on the region surrounding the selected SNP, if not the selected SNP itself. For each method, I calculate the average physical distance from the SNP under selection to the SNP demonstrating the strongest evidence for selection. This is an approximation for genetic distance and recognised as less accurate than calculating linkage disequilibrium with the selected SNP. Still, *Relate* and F_{ST} again demonstrate the highest accuracy in terms of identifying regions surrounding selected SNPs, showing the shortest distance between the selected SNP and that with the strongest evidence of selection (given for selected acting in the African population; **Fig. 2.5;** other metapopulation accuracy calculations are shown in **Figs. S2.3-5**). Also, as in line with previous analysis, accuracy remains higher when selection is further back in time and in the African population (**Fig. 2.5; Figs. S2.3-5**).

I therefore conclude that *Relate* and F_{ST} are most suitable for identifying selected SNPs (under SSV at these timepoints) in comparison to the haplotype-based methods tested here. SSV occurs on multiple haplotype backgrounds, especially when selection is acting on SNPs long after their emergence in a population, and selected SNPs are therefore found in significantly variable haplotypes. This lack of haplotype homozygosity reduces the power of haplotype-based methods to identify such selection. However, SSV still results in allele frequency differentiation (as identified by F_{ST}) and an unusually rapid spread of the selected SNP through the population (as identified by *Relate*). Moreover, *Relate* evaluates the probability of positive selection according to the history of each locus (in theory, complete and accurate history but, in reality, only inferred to the point of the common ancestor of all sampled populations), and therefore integrates more fully the complex evolutionary patterns compared to summary statistics, which prove important when considering the subtle signatures of SSV.

In summary, using the tails of the neutral distributions of *Relate* and F_{ST} results in the highest percentage of selected SNPs with evidence of selection and, whilst they are not able to accurately identify the exact selected SNP, they show moderate accuracy in suggesting the region containing the selected SNP.

2.4.2.1. The Effect of Frequency

Given that *Relate* and F_{ST} have the highest power to identify selected SNPs as those with evidence of selection, I now ask how the derived allele frequency (DAF) of the selected SNP may limit the power of these methods *i.e.,* at which DAF does accuracy appear to significantly drop. This is under the assumption that power to identify selected SNPs is highest at the highest DAF values, since these are the SNPs that likely show the most extreme allele frequency differentiation and trajectory through time (which are signatures of positive selection identified by F_{ST} and *Relate*, respectively). I use the previously calculated number of selected SNPs that are identified as showing evidence of selection (according to the 5% tail of the neutral distribution for either *Relate* or F_{ST}), conditioning on the DAF of the selected SNP, to compare how the proportion of selected SNPs identified as showing evidence of positive selection varies over different DAF values (**Fig 2.6**).

As expected, I observe the highest proportion of selected SNPs identified as showing evidence of selection, according to either *Relate* or F_{ST}, at the highest DAF values (shown for selection acting at 40kya in the African metapopulation in **Fig. 2.6,** shown for all other metapopulations in **Fig. S2.6**). According to *Relate*, the proportion of selected SNPs

identified as showing evidence of positive selection is at 49.53% when considering selected SNPs of all DAF values (for selection acting at 40kya in the African metapopulation). The proportion of correctly identified selected SNPs is higher than this baseline proportion when the DAF of the selected SNP is over 0.5 (**Fig. 2.6**). For F_{ST}, the proportion of selected SNPs identified as showing evidence of positive selection is at ~60-67% (depending on the cross-population comparison, for selection acting at 40kya in the African metapopulation) when considering selected SNPs of all DAF values. Here, however, the proportion of identified selected SNPs is higher when the DAF is over 0.4 (**Fig. 2.6**). I therefore conclude that the previously demonstrated accuracies can only be expected for these given DAF values or higher, and show that the accuracy drops significantly when considering positive selection acting on SNPs of lower DAF. I also conclude that F_{ST} is a little more robust to DAF variation, in comparison to *Relate*. Still, there are very few cases of low DAF (<20%) given that the simulations condition on the tagged variant being at 10% frequency or higher, and these results may therefore be noisy at lower DAF bins.

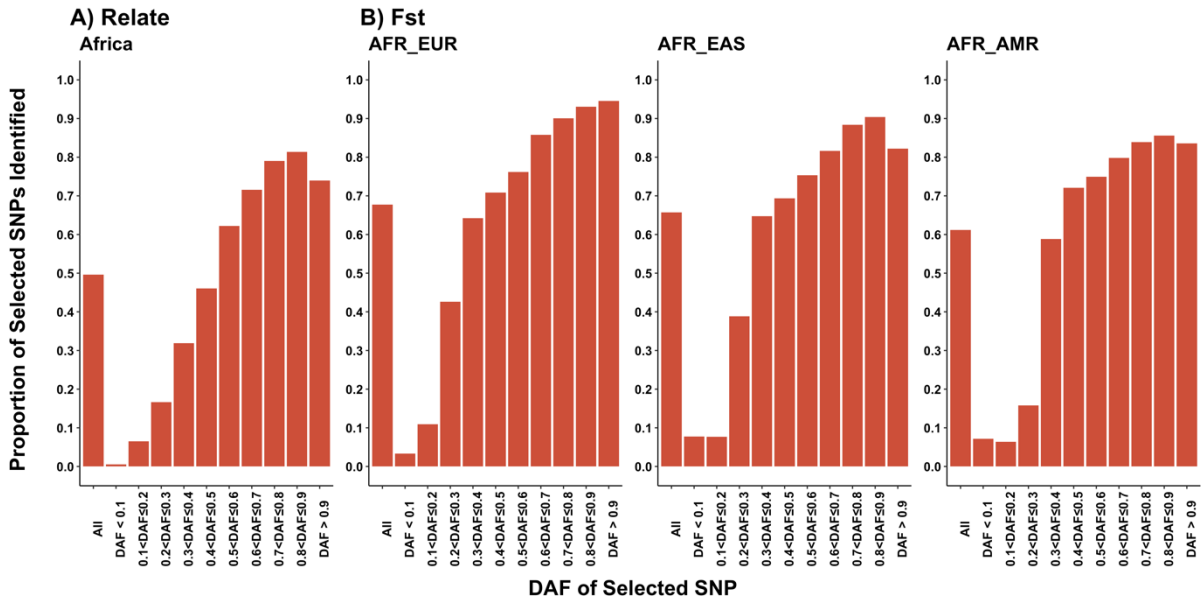


Fig. 2.6: Selected SNPs identified as under selection according to derived allele frequency. *The proportion of selected SNPs identified as under selection, partitioned by the DAF of the selected SNP. Given for selected SNPs identified according to the 5% tail of the neutral distributions of Relate (A) and F_{ST} (B; given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America). Shown for selection acting at 40kya for the African population.*

2.4.3. Identifying Polygenic Selection

2.4.3.1. Accuracy of the SUMSTAT Method

I now evaluate the accuracy of each of the methods to identify polygenic selection acting on gene sets using the *SUMSTAT* framework (Daub et al. 2013). These gene sets are built from either 10, 20, 40 or 60 simulated gene regions, where each gene region has a single SNP under positive selection (where all selection acting on a gene set is initiated at the same time in the same metapopulation). I identify gene sets as showing evidence for polygenic adaptation if the *SUMSTAT* value of that gene set (the sum of the most extreme outlier values for each statistic, see **Section 2.3.3.1**) falls in the 5% tail of the empirical neutral distribution for the *SUMSTAT* summed values integrating the respective statistic.

Reflecting the analysis of power to identify monogenic selection, I first observe remarkably high accuracy when using the F_{ST} method across all timepoints (shown for selection acting in the African population in **Fig. 2.7**; all other metapopulation analysis given in **Fig. S2.10**), and when using *Relate* method (shown for selection acting in the African population in **Fig. 2.7**; all other metapopulation analysis given in **Fig. S2.7**). However, the power of *Relate* is high only when selection is initiated at 40kya. Hence, I again recommend the use of these two methods to identify selection, but caution *Relate* loses power when selection is acting more recently across a gene set.

I also observe the general trend that increasing the size of the gene set increases the power to identify polygenic adaptation (the *Relate* and F_{ST} results summarised in **Fig. 2.7**, all methods across all metapopulations given in **Fig. S2.7-12**). Since gene set methods effectively combine signatures from multiple genes, here as a sum, it is thus expected that increased numbers of genes under selection increases the accuracy of these methods.

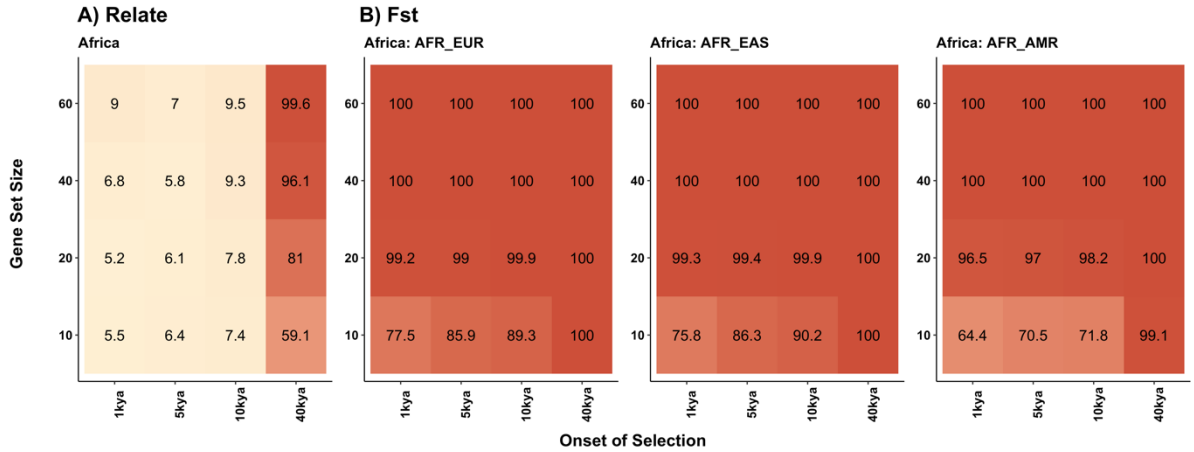


Fig. 2.7: Gene sets identified as under polygenic selection. *The percentage of gene sets identified as being under selection (according to the 5% tail of the neutral distributions) using the SUMSTAT method integrating Relate (A) and F_{ST} (B; given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America). Shown for selection acting at 40kya for the African population.*

I now evaluate how sensitive *Relate* and F_{ST} are to identifying selected SNPs under different selection coefficients, repeating the prior analysis but conditioning on the selection coefficients of all selected SNPs in a gene set.

I observe that *Relate* is more sensitive to the selection coefficients of the selected SNPs compared to F_{ST} (when selection is acting at 40kya in the African metapopulation; **Fig. 2.8**, all other metapopulation analysis for selection at 40kya given in **Fig. S2.13-17**). For example, for selection acting in the African metapopulation at 40kya, *Relate* drops from identifying 94.2% of polygenic adaptation gene sets with size 10 when the selection coefficients are between 0.005 and 0.004 to only identifying 9.3% when the selection coefficients are between 0.001 and 0.002 (when selection coefficients are uniformly distributed, *Relate* identifies 59.1% of gene sets as showing evidence of polygenic selection, see **Fig. 2.7**). Comparatively, for selection acting in the African metapopulation at 40kya, F_{ST} identifies 100% of polygenic adaptation gene sets with size 10 when the selection coefficients are between 0.005 and 0.004 and identifies 99.1% when the selection coefficients are between 0.001 and 0.002 (when selection coefficients are uniformly distributed, F_{ST} identifies 100% of gene sets as showing evidence of polygenic selection, see **Fig. 2.7**). Hence, I caution that *Relate* is considerably less accurate when selection is weaker across a gene set.

Indeed, F_{ST} within the *SUMSTAT* framework is almost always at an accuracy of 100%, implying that this may be a trivial evaluation of $F_{ST}'s$ power. Instead, the power of F_{ST} within a gene set could be more informatively evaluated at lower selection coefficients or gene set sizes (given that the broad pattern of decreased accuracy at lower gene set sizes and selection coefficients remains).

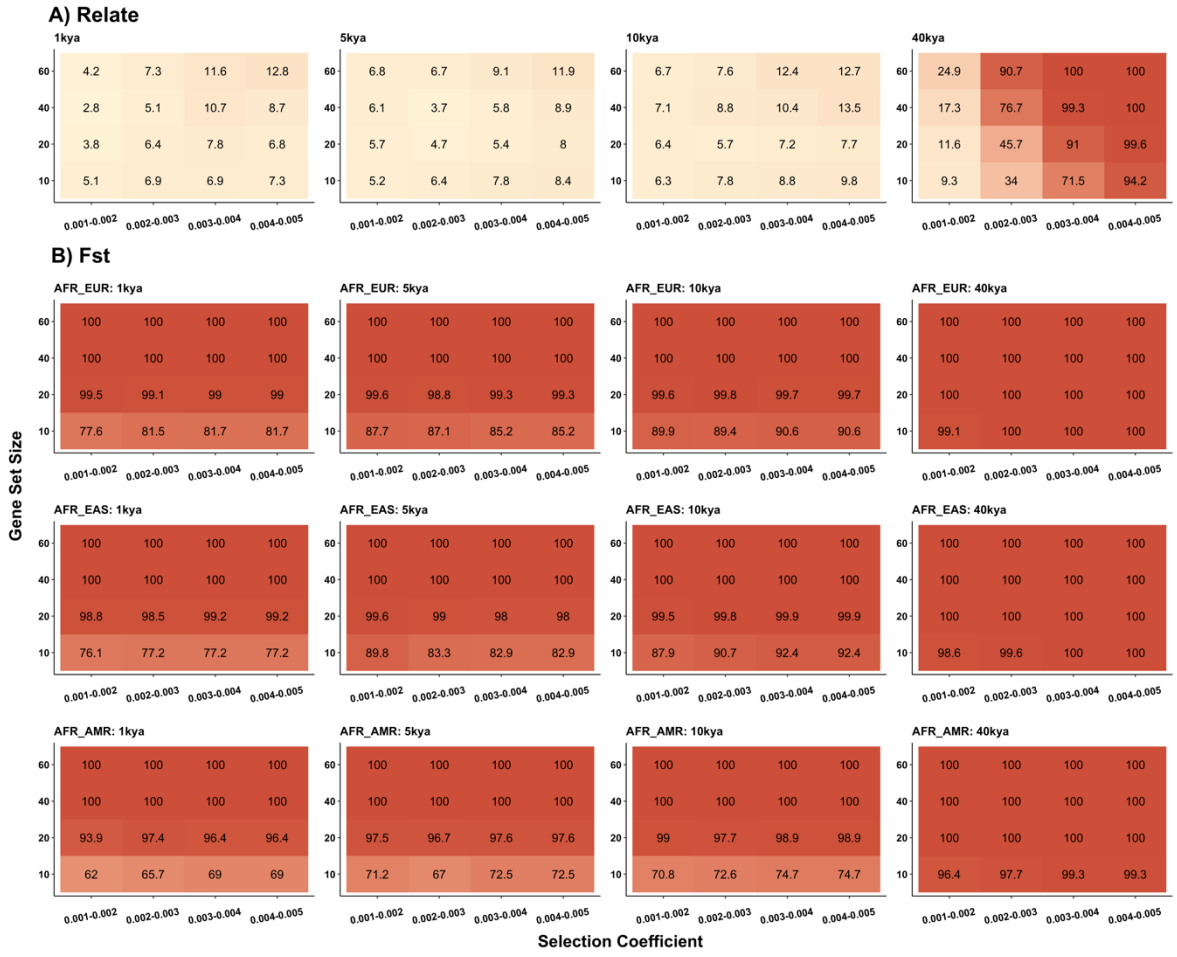


Fig. 2.8: Gene sets identified as under polygenic selectiona according to selection coefficient. *The percentage of gene sets identified as being under selection (according to the 5% tail of the neutral distributions) using the SUMSTAT method integrating Relate (A) and F_{ST} (B; given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America), partitioned by selection coefficient of the tagged* variant *(given for timepoints of selection of 1kya, 5kya, 10kya, 40kya). Shown for selection acting on the African population.*

2.4.3.2. Gene Sets of Both Neutral and Selected Genes

Finally, I consider the more realistic case where not all genes within a functional-related gene set evolve under the same selection, due to pleiotropy or other genomic constraints (Wagner and Zhang 2011; Fraïsse et al. 2019). Hence, I now evaluate the ability of *Relate* and F_{ST}, the most promising gene set methods, to identify gene sets as under polygenic selection when not all genes within a gene set experience selection. To do so, I vary the proportion of gene regions under positive selection, conditioning on only 20%, 40%, 60% and 80% of genes within a gene set as under selection.

As expected, I observe the highest accuracy when identifying selection on larger gene sets with the highest proportion of selected genes. When using *Relate*, and for selection starting at 40kya, the most marked increase in accuracy is when the proportion of

selected genes in the gene set is over 60%, and is particularly high when the gene sets are larger than 40 (**Fig. 2.9**, **Fig S2.18**). F_{ST} shows substantially higher accuracy at more recent timepoints, with gene sets containing only 40% of genes under selection showing impressive accuracy (given that gene sets are larger than 40 genes, **Fig. 2.9**, **Fig S2.4.19-22**).

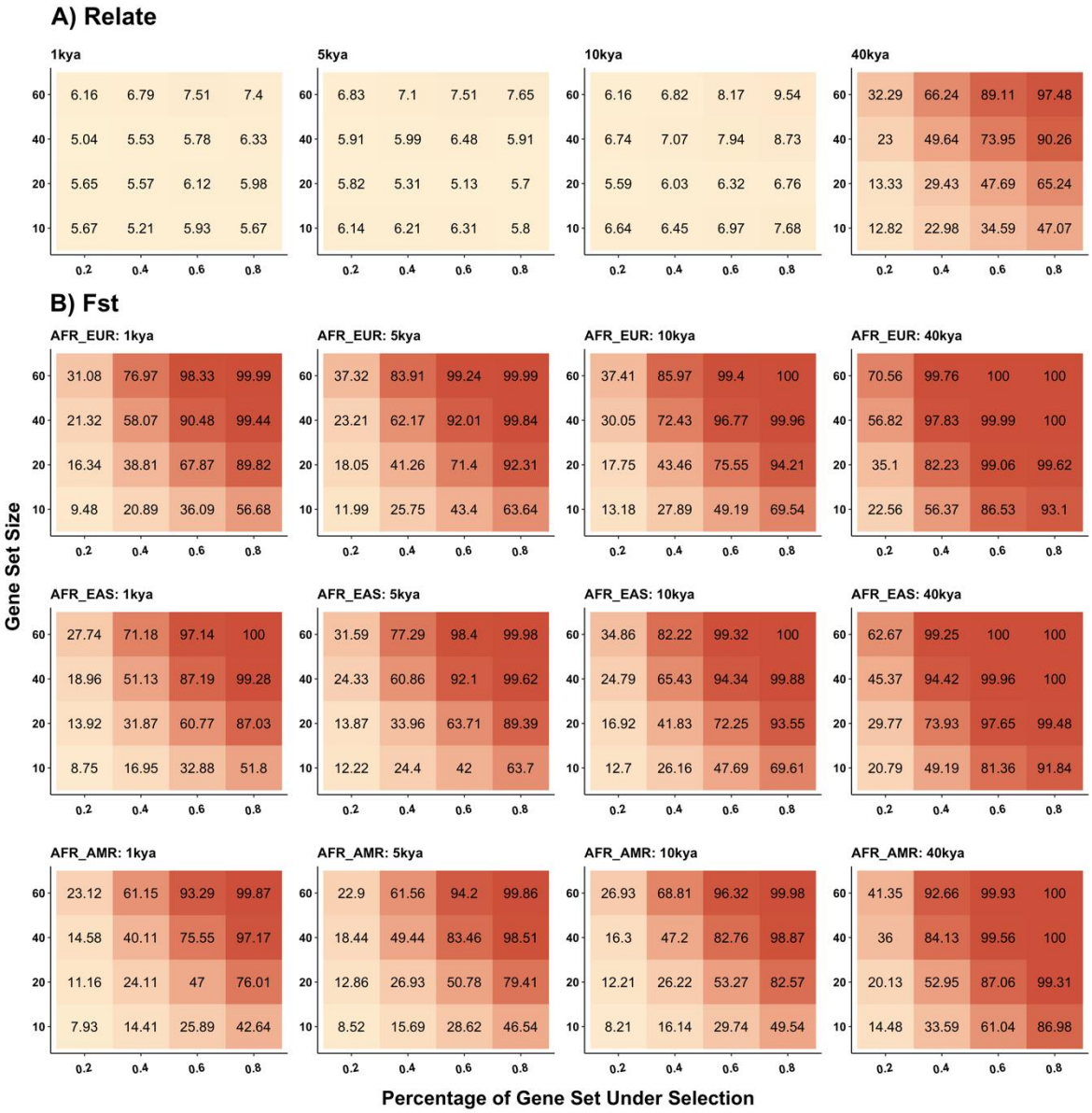


Fig. 2.9: Gene sets with varying proportions of genes under selection identified as under polygenic selection. *The percentage of gene sets identified as being under selection according to the SUMSTAT method, according to the proportion of the gene set under selection and given for timepoints of selection of 1kya, 5kya, 10kya, 40kya. Panel A corresponds to the SUMSTAT method integrating Relate; panel B corresponds to the SUMSTAT method* integrating F_{ST} *(given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America).*

Hence, I conclude that, when using *Relate* and F_{ST} within the *SUMSTAT* framework, power is one again highest for selection furthest back in time and for selection acting in the African metapopulation, as well as for larger gene sets with the highest proportions of genes under selection. Whilst I do recommend the use of *Relate* in the *SUMSTAT* framework, I caution that it is more sensitive to selection occurring at different timepoints and at different strengths in comparison to F_{ST}. It is likely that this is a product of the slightly higher power of F_{ST} to identify selected SNPs in individual gene regions as showing evidence of selection (**Fig. 2.4**) in comparison to *Relate* (**Fig. 2.3**).

2.5. Discussion

Common methods to identify positive selection are designed to identify related but subtly different signatures of selection. Summary statistics focus on identifying the signatures of allele frequency differentiation between populations, extended haplotype length or changes to the site frequency spectrum when compared with neutral expectations (Tajima 1989; Voight et al. 2006; Sabeti et al. 2007; Bhatia et al. 2013; Excoffier et al. 2013; Ferrer-Admetlla et al. 2014; Szpiech et al. 2021). More recent methods also focus on identifying selection events via the allele trajectory through time, inferred from reconstructed genealogies of loci across the genome (where such genealogies can now be inferred using the genomes of thousands of individuals; (Field et al. 2016; Kelleher et al. 2019; Speidel et al. 2019)).

However, the extent of these signatures or patterns of positive selection, and more importantly how accurately they are able to be drawn away from the neutral background of the genome, can vary wildly according to the exact dynamics of a selection event. The methods which aim to identify positive selection therefore present different accuracies according to relevant parameters such as the timepoint of selection, the demographic history of the population or the number and nature of alleles under selection. Forward simulation programmes, such as SLiM (Haller and Messer 2019), can explicitly model positive selection under these varied dynamics, and therefore facilitate the testing of these methods under different selective scenarios. In particular, SLiM is highly scriptable and therefore allows sophisticated customisation, well suited to modelling complicated genetic or selection scenarios, and is highly efficient, allowing the high numbers of simulations to be run in relatively little time (Haller and Messer 2019).

I designed a novel simulation framework using SLiM to test the accuracy of different methods to identify local adaptation in modern humans. I included selection events beginning at four timepoints in modern human history, two of which likely presented novel selective pressures to human populations: 10kya (approximate date of the Neolithic transition from a hunter-gathering to agricultural lifestyle (Latham 2013)) and 40kya (approximate date of major migrations to Eurasia (Seguin-Orlando et al. 2014)). The development from hunter-gatherer societies to those based on agriculture brought with it large changes to the human diet, as well as significant increases in population density resulting in increased risk of communicable diseases and zoonotic pathogens. Major migrations into varied Eurasian environments (and beyond) from approximately 40kya exposed colonising populations to novel temperatures and altitudes, as well as, in some cases, significantly altering the dietary composition and pathogen risk (Yi et al. 2010; Fumagalli et al. 2015; Mathieson et al. 2015a; White et al. 2015; Key et al. 2018). Hence, local adaptation events in modern humans are hypothesised to be driven in various populations at these timepoints. The remaining timepoints, 1kya and 5kya, were chosen to test the limits of all methods in identifying very recent selection.

The simulation framework also explicitly models selection that occurs on already segregating genetic diversity, or SSV. This mode of selection has been suggested to underlie some instances of local adaptation in modern humans but often remains elusive in the genome (Hermisson and Pennings 2005; Prezeworski et al. 2005; Schrider and Kern 2016, 2017). SSV has been particularly suggested to be the dominant mode of adaptation for populations rapidly encountering novel environments (Schrider and Kern 2016, 2017; Hermisson and Pennings 2017), since low frequency alleles maintained in an expanding population would have mediated faster adaptation than *de novo* mutations (de Filippo et al. 2016; Hermisson and Pennings 2017).

Finally, I extended the framework to consider polygenic selection, grouping the simulated genomic regions into polygenic adaptation gene sets. Polygenic selection has been suggested to be prevalent in human evolutionary history given that it likely underpins the adaptation of complex traits, such as those relating to immunity, diet and metabolism (Pritchard et al. 2010). However, the weak and varied signatures of selection spread over different genomic regions, characteristic of polygenic selection, are also often difficult to confidently identify (but not impossible; (Fumagalli et al. 2011; White et al. 2015; Nédélec et al. 2016; Berg, Harpak, et al. 2019; Berg, Zhang, et al. 2019; Roca-Umbert et al. 2022)). Hence, by designing the simulations to model both SSV and polygenic adaptation, and using selection coefficients that represent weak selection ($0.001 < s < 0.005$), I am able to evaluate the methods with the highest power to identify the local adaptation events that are likely present, but remain somewhat elusive, in our history.

From these simulations, I identify two methods which demonstrate the highest power in identifying local adaptation in modern humans ($F_{ST}$and the $Relate$) and compare them to the weaker performing haplotype-based methods. I specify that the power of $F_{ST}$and the $Relate$ is primarily demonstrated only in the cases of older selection, that occurring more than 10kya, and is particularly observed in those populations with high N_e (and by consequence, having higher genetic diversity). Despite their high power to identify strong signatures of selection (Voight et al. 2006; Sabeti et al. 2007; Ferrer-Admetlla et al. 2014; Huerta-Sánchez et al. 2014; Szpiech et al. 2021), all haplotype methods used here have low power to identify weak selection on standing variation. The diverse genetic backgrounds of which the adaptive allele is on, and the weak selection simulated, result in the absence of long and uniform haplotypes within a population. Hence, it is expected that the power of haplotype-based methods is so low under this particular selective scenario.

2.5.1. F_{ST}

F_{ST} has the highest power to identify selected SNPs as those with evidence of selection over each timepoint of selection, and appears to be more suited to identifying selection across a range of metapopulations in comparison to $Relate$ (which has larger differences in power across the metapopulations). The increased power of F_{ST} likely comes from its cross-population approach; by comparing allele frequency across populations, this method is able to more accurately identify positive selection isolated to one population. Whilst $Relate$ identifies SNPs with an unusual inferred trajectory through time, it does not compare between populations and therefore cannot integrate this comparison into its assessment of the likelihood of allele trajectory.

2.5.2. *Relate*

Relate integrates the entire inferred history of the allele to calculate the probability of a variant having its inferred spread through a population. Since I use the tails of the neutral distribution to identify selected sites, here I explicitly identify those sites with an unusually rapid spread through a population relative to all other neutral sites in that population. It has been previously suggested that methods that more fully integrate the history of an allele, such as these genealogical methods, are more suited to identifying weaker selection, as they do not depend on a single strong signature in modern genomes, and do not remove key information (which may indicate selection) when summarising complex patterns into a single statistic. Indeed, I show that this is a powerful approach when identifying weak SSV when the onset of selection is older than 10kya.

However, the power of *Relate* remains low when identifying selection at 5kya or 1kya. I believe that this, in part, is due to *Relate* inferring the probability of selection of an allele since the appearance of the mutation. This means that the underlying assumption of selection is that it occurs immediately on the birth of a mutation, assuming SDN rather than SSV (as simulated here). When selection occurs further back in time, the birth of the mutation and onset of selection are likely to be significantly closer, perhaps indistinguishable, compared to the long period of neutrality a variant may have before a selection onset at 5kya or 1kya. Hence, signatures of SSV may be lost within the trajectory of the allele over its lifetime. The output of *Relate* can be customised, however, to output the probability of positive selection from a specified number of generations ago. Rather than using the raw $-log10pvalue$ to evaluate the evidence of selection at this timepoint (which will be adversely affected by the lower power to identify selection at more recent timepoints), I suggest using the tail of the empirical background distribution of probabilities inferred from the same timepoint to identify the likelihood of selection on a SNP. In theory, this would identify the SNPs along the genome with the most unusual trajectory from the given timepoint. Still, this relies on assumptions of the timing of selection; identifying SNPs with evidence of positive selection whilst simultaneously suggesting the timing of the selection is considerably more sophisticated (but has been attempted; (Stern et al. 2019)).

The power of *Relate* also differs amongst metapopulations of different demographic histories, with the simulated African metapopulation having markedly higher power in comparison to the other metapopulations. The reduction in power to identify positive selection in the European, East Asian and American populations, biased by the simulation design but remaining informative (see **Section 2.3.3.1**), is likely due to the decreased N_e and consequent decreased genetic diversity (not independent from the bottlenecks in their demographic histories). In turn, this reduces the number of lineages in the genealogies of each selected SNP. Since the *Relate* test for selection conditions on these lineages (Speidel et al. 2019), a reduction in lineages likely reduces the power to identify selection. As alluded to in **Section 2.5.1**, a cross-population approach (comparing the trajectories between populations) may reduce the power imbalances between populations.

Further, I demonstrate that whilst *Relate* is vulnerable to low sample sizes, the number of selected SNPs identified can be increased by using the tails of the empirical neutral distribution rather than the raw probability output, as previously used (Speidel et al. 2019). Hence, I suggest that this is the most powerful way of using this method to identify

positive selection, and recommend this method particularly when identifying selection in populations with low N_e.

2.5.3. F_{ST} and *Relate* in Identifying Polygenic Adaptation

The relatively high accuracy of F_{ST} and *Relate* is also demonstrated when considering local selection on a polygenic adaptation gene set, as evaluated using *SUMSTAT* (Daub et al. 2013). As expected, I show that the accuracy of these methods increases with larger gene set size, since many small shifts in the test statistic are needed to significantly shift the sum value to the tail of the empirical neutral distribution. When using *Relate* to identify polygenic selection according to the *SUMSTAT* framework, gene sets should have approximately 60% or more of their genes as under selection for appreciable power. This proportion can drop to approximately 40% when using F_{ST}. I therefore suggest that, even though there is higher power to identify polygenic selection on larger gene sets, it is the proportion of genes under selection that is most important in governing the power of the *SUMSTAT* approach. Hence, the *SUMSTAT* approach is limited to identifying polygenic selection on gene sets where the majority of genes are responding to selective pressures.

I also show that the power of F_{ST} to identify polygenic selection remains high even when selection acting on the gene set is weak ($0.001 < s < 0.002$; (Turchin et al. 2012)). In contrast, the power of *Relate* to identify polygenic selection on gene sets appears to be considerably more sensitive to the selection coefficients than F_{ST}. When selection is this weak, the allele trajectory is likely too similar to that expected under neutral drift and *Relate* is unable to draw this signature out from the rest of the genome. The programme's underlying assumption of SDN, or inference of the probability of selection of variant since its appearance, likely further hinders the identification of weak selection; the effect of weak selection on the trajectory of a variant segregating neutrally for some time is not strong enough to characterise the entire lifetime of a variant as being under selection (especially if the time of the variant evolving under neutrality is very long). Hence, I again suggest that the current assumptions of *Relate* place some limitations on its ability to identify SSV.

2.5.4. Limitations and Future Directions

I demonstrate the power of F_{ST} and the *Relate* method to identify weak positive selection acting on standing variation as recent as 10,000 years ago, but recognise that the inferences from these simulations are limited by a few key factors. I simplify the mutational landscape to only one positive mutation and do not consider the effects of pleiotropy and epistasis, which may limit the response of a genomic region under selection (and is particularly relevant when considering polygenic selection on a gene set). I also use a relatively simply demographic model, and caution that the demographic history of each metapopulation does not accurately represent that of all populations within that region. Finally, since weak to moderate selection coefficients are used (Turchin et al. 2012), the estimates of power of these methods are conservative, and may indeed be significantly higher when identifying stronger SSV (and certainly under strong SDN).

I show that the present methodological toolkit is well-equipped to identify selection surrounding the timepoints of major migratory events into non-African environments (~40kya) and large cultural changes (*i.e.,* the Neolithic revolution, ~10kya). However, it

is clear that issues of accuracy remain when identifying more recent selection. Mass disease, continued dietary change and even temperature fluctuations (Allentoft et al. 2015; Mathieson et al. 2015a; Demény et al. 2019) have continued to exert significant selective pressure on human populations in the last 5000 years, but there are few, if any, methods that are reliably able to identify that but the very strongest selection (Rees et al. 2020). It is likely that the use of ancient genomes will significantly help in identifying this recent selection, since they can provide direct insight into past allele frequency. Given enough samples, ancient DNA can thus pinpoint the timing of rapid allele frequency change and inform inferences on the timing of selection events.

Still, I highlight here the promise of tree-based statistics. Tree-based methods to identify positive selection are in their relative infancy, but here I already demonstrate their ability to identify weak selection as recent as 10kya. Methods that are similar or derived from *Relate* may prove more powerful if the underlying assumption of SDN can be removed, or they are able to integrate assumptions on the onset of selection to identify unusual trajectories, relative to the empirical neutral background, following this point. By integrating ancient DNA into these methods, either to constrain the inferred tree as in (Wohns et al. 2022) or to suggest the onset of selection, these methods will likely continue to improve.

It remains that by considering the entire inferred history of an allele, and hence avoiding the collapse of complex evolutionary patterns into a singular value, the field has a powerful way to identify subtle signatures of selection, and one that will likely progress rapidly. Indeed, current packages to identify the signatures of positive selection on tree sequences are already highly customisable (Kelleher et al. 2019) and may well be developed under recent, SSV assumptions. Ultimately, focusing on allele trajectory over time (as inferred by tree-based methods), as well as allele frequency differentiation, are the most promising avenues to identify weak SSV, since the rise in allele frequency characterises all positive selection.

Finally, I suggest that many studies may benefit from integrating simulations of a similar design: those that explicitly model target populations' demographic history (more so than done here) to identify the methods with the highest power to detect positive selection at hypothesised timepoints. Naturally, this requires a clear, if not approximate, understanding of the study population's demographic history, and I recognise the barriers in accurately modelling these. However, if such demographic histories can be confidently inferred, simulations such as these allow a considerably more informed view on which methods are most powerful in identifying hypothesised selection events and may aid the identification of previously elusive signatures of positive selection.

2.6. Conclusions

Using a novel simulation framework that models weak selection on standing variation in individual metapopulations, I demonstrate allele-differentiation and tree-recording methods as having the highest power to identify the genetic signatures of local adaptation in modern humans up to 40,000 years ago. These findings extend to polygenic selection using a gene set method, and I observe a significant drop in accuracy when selection starts less than 10,000 years ago.

Chapter 3: Signatures of Adaptation to Micronutrient-Associated Genes in Modern Humans

3.1. Overview

Trace minerals, macrominerals and vitamins are essential dietary components for human health, with pathologies occurring below or above their narrow, recommended range (Tako 2019; De Groote et al. 2021). These micronutrients accumulate in the diet according to the local soils, which affect the micronutrient composition in plants and their consumers (von Wandruszka 2006; Hurst et al. 2013; De Groote et al. 2021). Indeed, micronutrient soil levels are highly variable across the globe, with both local pockets and wide-spread regions of soil that are deficient or toxic for any given micronutrient (Karimov et al. 2009; Hurst et al. 2013; Hengl et al. 2017; Nell and van Huyssteen 2018; De Groote et al. 2021). Hence, human populations occupying different environments are exposed to varying levels of these essential micronutrients, which may act as a local selective pressure to drive adaptations in the genes involved in their metabolism, uptake or transport.

In this study, I use methods based on allele frequency differentiation and recently developed tree-recording methods (Weir and Cockerham 1984; Speidel et al. 2019) to infer signatures of natural selection across 276 micronutrient-associated genes, linked to the uptake, metabolism or regulation of 13 micronutrients in 40 diverse modern human populations (Bergström et al. 2020). I show that such signatures are present across many global populations and micronutrient categories, and the strongest signatures of natural selection recapitulate known geology and endemic deficiencies in modern human populations. I do not see evidence for micronutrient-associated adaptation being mediated by polygenic selection and suggest that micronutrient-associated adaptation is largely mediated by monogenic or oligogenic selection. Finally, I propose the micronutrient-associated gene sets and individual micronutrient-associated genes with the strongest evidence of positive selection in global populations.

3.2. Background

Diet is a dominant selective pressure across all organisms, driving adaptation to uptake, regulate and metabolise key dietary components (Perry et al. 2007; Drouin et al. 2011; Li and Zhang 2014; Wu 2022). In modern humans, diets are highly diverse across the globe (Fumagalli et al. 2015; Fan et al. 2016; Rees et al. 2020) and hence exert differential selective pressures amongst different populations. This can result in local adaptation: genetic adaptation in specific populations, or groups of populations, that results in a local adaptive phenotype.

Indeed, dietary drivers of local adaptation are well-recorded in modern humans, driven either by environmental or cultural differences in the local diet amongst populations. These adaptations may be driven by novel sources of nutrition (such as lactase (Tishkoff et al. 2007a)), toxic substances (such as alcohol or toxic metals (Osier et al. 2002; Han et al. 2007; Schlebusch et al. 2015), differing proportions of key macronutrients (such as

fatty acids or starch (Perry et al. 2007; Fumagalli et al. 2015)) or nutritional abundance (such as highly variable dietary conditions (Minster et al. 2016)).

3.2.1. Micronutrients in the Human Diet

Micronutrients, which include trace minerals, macrominerals and vitamins, are a key component of the human diet and a likely selective pressure in modern humans. With the exception of vitamin D, micronutrients cannot be synthesised by the human body, and therefore must be absorbed via the diet. Moreover, the healthy range of micronutrient levels in the diet is very narrow (Renwick 2006), particularly for trace minerals (which are recommended at daily levels of 50micrograms to 18milligrams (Mertz 1981)). For example, the recommended levels of the trace mineral level zinc are 8-11mg (De Groote et al. 2021) with deficiency being induced when consuming daily levels as high as 5mg (Prasad 2013). Macrominerals, whilst recommended in higher amounts than trace minerals (daily recommendations >100mg (Tako 2019)), are still required at much narrower levels than macronutrients.

Micronutrient deficiencies are relatively prevalent (Caballero 2002; Xia et al. 2005; Renwick 2006; Shenkin 2006; Bhutta and Salam 2012; Biban and Lichiardopol 2017; De Groote et al. 2021), with iron and iodine-deficiencies most common across the globe and estimated to affect approximately 25% of the world's population (Bhutta and Salam 2012; Bailey et al. 2015). Micronutrient toxicity is comparatively rarer, and is usually a result of over-supplementation (Renwick 2006; Pike and Zlotkin 2019). Common adverse effects resulting from toxicity include gastrointestinal distress, nausea, vomiting and diarrhoea, as well as increased interaction with non-essential chemicals (Peraza et al. 1998; Pike and Zlotkin 2019).

Micronutrient deficiencies vary in their exact pathology, but all increase the risk of various metabolic, infectious and respiratory diseases, as well as often impairing mental and physical development (Caballero 2002; Tulchinsky 2010; Prasad 2013; Biban and Lichiardopol 2017). Common pathologies resulting from trace mineral and macromineral deficiency include anaemia (resulting from iron deficiency across many populations (Caballero 2002)), goitre (iodine-deficiency, commonly reported in mountainous or some forest environments (Dormitzer et al. 1989; Niepomniszcze et al. 2009; Biban and Lichiardopol 2017)), and the heart and bone diseases Keshan and Kashin-Beck diseases (linked to selenium-deficiency, endemic to selenium-deficient areas of China (Moreno-Reyes et al. 2001; Xia et al. 2005)).

3.2.2. Micronutrients in Human Local Adaptation

The essentiality of micronutrients in the human diet, alongside serious pathologies that accompany deficient levels, mean that dietary micronutrients are a strong candidate selective pressure within human evolution and hence may drive local adaptation. Moreover, this selective pressure has likely been differentially exerted over populations. Whilst today omission of key food groups can result in micronutrient deficiencies, for much of human history the levels of dietary micronutrients was determined by those available in the local consumed animal and plant products. In turn, this was heavily influenced by the geology and micronutrient composition of the local soil (Sillanpaeae 1982; Hurst et al. 2013; Prasad 2013). Since soil levels can vary widely even between proximal localities, on the level of the individual populations rather than across continents (*e.g.,* extremely selenium, zinc and iodine-deficient soils in areas of China,

Ethiopia and Central Africa, respectively; Cifor, 2006; De Groote et al., 2021; Dhaliwal et al., 2019; Dormitzer et al., 1989; Hengl et al., 2017; Xia et al., 2005) this may have resulted in relatively fine-scale local adaptation to micronutrient uptake, regulation or metabolism.

Indeed, selenium-deficient soil in East Asia has been associated with adaptation in selenium-associated genes, particularly in *DIO2, SelS, GPx1, CELF1* and *SEPHS2* (White et al. 2015). A correlation between zinc levels in soil and crops in East Asia with a particular haplotype of the zinc-associated gene *SLC30A9* has also been inferred (Zhang et al. 2015a). Finally, iodine-deficient soil in rainforest environments have been suggested to drive potential signatures of positive selection in *TRIP4* and *IYD* genes in the Biaka population (Herráez et al. 2009). This adaptive scenario is supported by the lower incidence of goitre in rainforest pygmy populations compared to the neighbouring Bantu populations (42.9% compared to 9.1%; (Dormitzer et al. 1989).

Dietary micronutrient levels are not only affected by the local soil, but also by the exact content of the diet, more closely tied to cultural differences and dietary evolution amongst populations. The rapid changes in the diet during the Neolithic revolution included a reduction in nutrient-rich animal products in favour of a cereal-based diet, dominated by staple crops and lacking key nutrients such as iron and calcium (Diamond 2002; Naugler 2008). Such recent dietary changes have also then been suggested to drive adaptation in iron and calcium-associated genes, namely *HFE* and *TRPV6* in European populations (although the former has also been suggested to be a result of allele-surfing: the geographic spread and increase in frequency of alleles during a range expansion that may mimic the signatures of positive selection (Akey et al. 2004; Distante et al. 2004; Ye et al. 2015; Peischl et al. 2016)). Agricultural practices born from the Neolithic revolution may also deplete soils of key micronutrients (Diamond 2002), and populations may then have also experienced increased micronutrient-associated stress, as a result of decreased micronutrients in the soil and therefore diet.

Hence, dietary micronutrient levels are not only a strong candidate selective pressure in human evolution, but one that may be exerted differentially amongst populations to ultimately result in local adaptation. This may be driven by the micronutrient content of global soils, cultural evolution of diet or the development of human agricultural practices (although, it is unclear the extent of micronutrient depletion farming would have imposed before very recent times).

Still, here three additional selective drivers are mentioned, unrelated to local soil or cultural evolution, suggested to be driving particular examples of micronutrient-associated adaptation in modern humans. The first is the degree of UV exposure experienced by a population. Whilst level of ingested calcium depends on the content of the diet, the extent of calcium absorption relies on adequate vitamin D levels, which in turn is produced on exposure of UV light (Carlberg 2022). Because of this, it has been suggested that the low UV levels in some populations, specifically northern European populations, instead act as the selective driver for calcium uptake (Mathieson and Terhorst 2022). The second proposed selective driver is the pathogen stress experienced by a population: reducing intracellular levels of zinc and iron has been suggested to be an adaptive response that starves pathogens of their essential micronutrients, thereby reducing the risk of serious infection (Engelken et al. 2014; Pietrangelo 2015). Finally, ambient temperature has been suggested to have driven iron-related adaptation in European populations, as a connection between the

thermoregulatory role of iron and the colder temperatures experienced by populations within Europe compared to Africa (Heath et al. 2016a).

It is therefore clear that a multitude of questions remain surrounding micronutrient-associated adaptation in humans. This includes to what extent such proposed adaptation has played in human genomic variation (and in response to which micronutrients) and which exact selective pressures are most likely to be driving potential micronutrient-associated adaptation events. It is also unclear if micronutrient-associated adaptation may be polygenic in nature (as suggested by some studies (White et al. 2015; Zhang et al. 2015a)), and how this may vary across micronutrients.

At the time of writing, there has been no comprehensive study that investigates adaptation in modern humans across micronutrients and across global populations, thus limiting our knowledge to individual studies exploring specific micronutrient-associated adaptation. From these existing studies, it is not possible to fully evaluate the role dietary micronutrients have played in driving human genetic adaptation, and not able to compare the role of each micronutrient in such genetic adaptation. Moreover, many previous studies have been carried out in limited population cohorts and do not fully represent the geographic and genetic diversity of modern humans, and hence are not able to comprehensively evaluate the geographic distribution of potential adaptations. Finally, the current literature shows considerable bias towards particular micronutrients or genes, and there is little known about adaptation to the still-essential micronutrients, such as magnesium or phosphorus.

3.2.3. Study Overview

Here, I carry out a comprehensive study exploring selection in just under 300 genes associated with the uptake, metabolism or regulation of 13 trace minerals and macrominerals. I use simulation-informed methods (see **Chapter 2**) of allele-frequency-differentiation (F_{ST}; (Weir and Cockerham 1984)) and genealogical inference ($Relate$; (Speidel et al. 2019)) to identify instances of local adaptation across 40 genetically and geographically diverse modern human populations (Bergström et al. 2020).

I show that signatures of natural selection are present for many micronutrient-associated genes in many global populations, in some cases supported by known soil levels and dietary deficiencies in modern human populations. I find no evidence that selection acts over entire micronutrient gene-sets, and infer that adaptation is more likely oligogenic than polygenic in nature. I also identify the populations and the candidate genes with the strongest evidence of having undergone positive selection in response to micronutrient levels, and ultimately suggest that dietary micronutrients have played a role in shaping the genetic diversity of our species.

3.3. Methods

3.3.1. Micronutrient-Associated Gene Sets

I curate gene sets associated with the uptake, regulation and metabolism of 13 micronutrients: selenium, copper, iron, zinc, iodine, manganese, molybdenum, calcium, phosphorus, magnesium, sodium, chloride and potassium. This includes all trace minerals (N=7; selenium, copper, iron, zinc, iodine, manganese, molybdenum) and macrominerals (N=6; calcium, phosphorus, magnesium, sodium, chloride, potassium)

with the exception of fluoride and sulfur, which were omitted due to limited literature surrounding their functionally-associated genes in modern humans.

Gene sets (see **Table 3.1**) were manually created from relevant databases (*e.g.,* Human Metabolome Database (Wishart et al. 2007)) and a literature search using key terms including "human health", "metabolism", "adaptation" for each specified micronutrient. The literature used includes clinical studies, functional biochemical studies and studies identifying signatures of natural selection (see **Table S3.1**). Signatures of natural selection have only been identified in genes associated with selenium, zinc, iron, calcium and iodine, and such genes make up only a small proportion of the gene sets (see **Table S3.1)**. Hence, the ascertainment bias from this literature search is in this regard as minor.

In total, 276 micronutrient-associated genes (MA-genes) were identified, 263 of which are autosomal. After the filtering step that removes segments of the genome of low reliability (according to a positive mask, see **Section 3.3.3** (Bergström et al. 2020)), 269 genes remain. The micronutrient-associated gene sets vary in size (see **Table 3.1**), somewhat reflecting the number of genes associated with specific micronutrient uptake and metabolism, but also recognised as reflecting a bias of the available literature by micronutrient. This is considered during the following analysis, *e.g.,* how this may affect the proportions of genes showing signatures of positive selection in each gene set (see **Section 3.4.3**).

Notably, some genes are associated with multiple micronutrients (common overlaps are between selenium and iodine, sodium and potassium, and calcium and phosphorus; see **Table S3.1**). For some analyses, cut-down micronutrient gene sets where there exists no overlap are used, and each gene is only assigned to its most strongly associated micronutrient, according to the available literature (see **Table S3.1**).

***Table 3.1**: **The total number of genes used in this study associated with the uptake, metabolism or regulation of 13 micronutrients.** Number of genes for each MA-gene set given as a total ("Total Set"), the number of genes following the removal of any masked gene regions ("Post-mask") and when cutting down gene sets to remove any overlap, assigning each gene to its most supported associated micronutrient set ("Cut-down").*

		Number of Associated Genes		
	Micronutrient	**Total Set**	**Post-mask**	**Cut-down**
Trace Minerals	Selenium	61	59	59
	Copper	11	11	9
	Iron	44	44	44
	Zinc	46	45	42
	Iodine	18	18	14
	Manganese	7	7	4
	Molybdenum	5	5	5
Macrominerals	Calcium	23	21	17
	Phosphorus	16	16	14
	Magnesium	19	19	15
	Sodium	20	20	17
	Chloride	25	23	22
	Potassium	11	11	7

3.3.1.1. Distribution of Micronutrient-Associated Genes

The gene regions for each of the MA-genes were extracted from the *ensembl* database (Yates et al. 2020), and those which have overlapping gene regions or are less than 10kbp apart were identified (see **Table S3.2**). Any signatures of positive selection identified in these overlapping gene regions are treated as possible signatures for either gene region, rather than assigning to a single MA-gene. I verify that the MA-genes are, on average, randomly distributed along the human genome using *ChromoMap* (**Fig. S3.1**; (Anand and Rodriguez Lopez 2022)).

Since elevated allele frequencies can lead to false positives in selection scans (Buffalo and Coop 2020), I also evaluate whether the distribution of derived allele frequencies across micronutrient gene sets are significantly higher than the genomic background. For each micronutrient gene set, I sample the equivalent number of SNPs from chr1 of 22 Yoruba individuals (Bergström et al. 2020) and compare the distribution of allele frequencies between the SNPs in the gene set and this background distribution using a Mann-Whitney test.

Only four micronutrient gene sets have a significantly different allele frequency distribution compared to the genomic background (copper, magnesium, sodium and molybdenum; MW test, $p < 0.05$; see **Table S3.3**). The differences between the micronutrient set and genomic background are either negligible (*e.g.,* the mean allele frequency difference between the background and the magnesium and sodium gene sets) or the mean allele frequency is lower in the gene set than in the background distribution (*e.g.,* copper), so these differences are treated as irrelevant to this study. The remaining difference, the significantly higher allele frequency in the Molybdenum gene set (n=5) than in the background distribution, appears to be driven by the *GPHN* and *MOCS2* genes (possibly as a result of positive selection, as suggested in **Section 3.4.3**).

3.3.1.2. Generating Matched Neutral Gene Sets

A database of neutral gene regions matched to each MA-gene set is generated, accounting for the number, length and SNP density of genes within each set. For each MA-gene, I sample 1,500 gene regions of equivalent length from the human genome beginning at the starting genomic coordinate of a random human gene. I retain the 1,000 gene regions with the SNP densities closest to each associated MA-gene (SNP densities sampled from the genomes of Yoruba individuals (Bergström et al. 2020)). This results in a random set of gene regions (proxy MA-gene regions, now referred to as pMA-gene regions) which represent the genomic background, approximately matched in length and SNP density to the MA-genes.

The SNP density (number of SNPs above 5% in the Yoruba individuals) of seven MA-genes fall above the 95th percentile of the SNP density of their respective pMA-gene regions, where the 95th percentile is calculated from the cumulative frequency distribution (CDF). These seven MA-genes are thus noted as SNP-dense genes: *SELENOO* (selenium-associated), *EPAS1* (iron-associated; high SNP-density likely explained by its introgression from Denisovans (Huerta-Sánchez et al. 2014)), *MT1A* and *MT1F* (zinc-associated), *SCNN1D* (sodium and potassium-associated), *SLC8A1* (calcium-associated) and *CLCN7* (chloride-associated) (see **Table S3.4**). Still, these SNP-dense genes do not cluster by micronutrient. Calcium is the only MA-gene set with SNP densities, over the entire gene set, that are significantly shifted towards higher SNP density than the background (inferred from the distribution of CDF values; see **Table S3.5, Fig. S3.2**). This seems to be driven by a clustering of CDF values around 0.5-0.7, hence the deviation is not considered extreme. Moreover, there is no deviation across all other gene sets.

3.3.2. The Population Dataset

I use 929 full human genomes from the HGDP dataset (as published by (Bergström et al. 2020)), which encompasses 54 populations across Africa, the Middle-east, Europe, East Asia, Central-South Asia, Oceania and the Americas and represents a significant proportion of human ethnic and cultural diversity. Since low sample sizes can significantly reduce the power to identify the genomic signatures of positive selection (see **Chapter 2**; (Subramanian 2016; Serdar et al. 2021)), I aim to merge populations with sample sizes below 20 with their geographically closest populations. In these cases, the signatures of fine scale positive selection in response to extremely localised micronutrient soil levels may be lost, but this a necessary step to maintain adequate power to identify positive selection that may be shared across these geographically close populations.

Population analysis was carried out to verify that this criterion agreed in all cases with patterns of population differentiation. I calculated principal components (PCs; linear combinations of the initial SNP data) for each metapopulation using $plink$ (Purcell et al. 2007), having thinned for linkage disequilibrium (pruning r^2 values above 0.2) and using windows of 50kbp and window step size of 10bp (see **Fig. S3.3.3-9**). I also carried out clustering analyses on chr14, chosen due to its middling size in the genome, using the $admixture$ programme (Alexander et al. 2009) and the same linkage disequilibrium filtering for a varying number of clusters on the African, European, East Asian and American populations to aid population grouping (see **Fig. S3.3.10-13**). This analysis confirmed that grouping by geography agrees with population differentiation, with two exceptions (see below), and I hence group according to this criterion.

When grouped, the final dataset comprised of 913 individuals from 40 populations, of which 10 are a result of merging (see **Fig. 3.1, Table S3.6**). Two merged populations do not follow geography (Bantu-speaking population and the Xibo-Mongolian population), but instead reflect recent migrations (Bai et al. 2014; Patin et al. 2017; Hou et al. 2022). Two populations were removed from the analysis (Columbian, n=7; Cambodian, n=9) since they do not group naturally geographically or genetically. Despite their small sample size, the San population (n=6) was retained in the final dataset given their relatively distinct genetic variation.

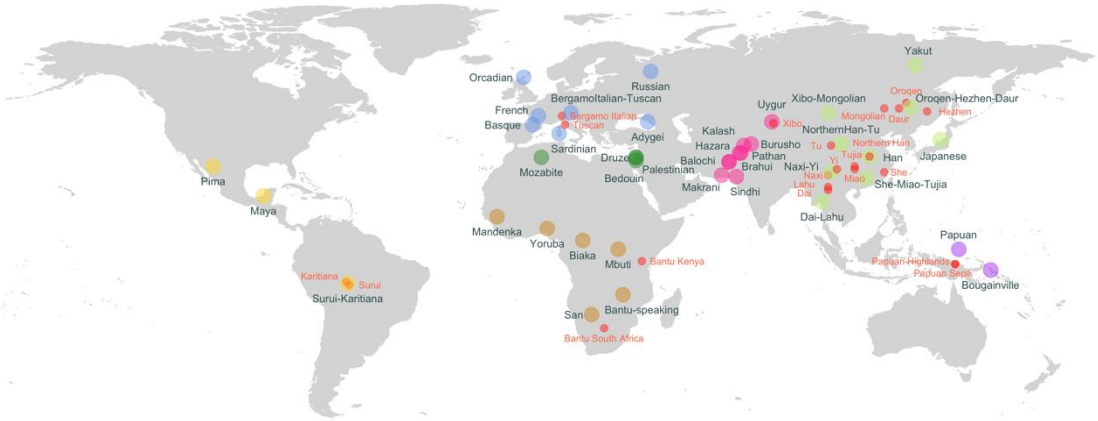


Figure 3.1: Map of the populations in this dataset. *The final populations, including merged populations, used in this study (large circles: dark orange = African; dark green = Middle-East; blue = European; pink = Central-South Asia; light green = East Asian; gold = American; purple = Oceania). Smaller red circles indicate the location of original populations merged together.*

3.3.3. Methods to Identify the Genomic Signatures of Positive Selection

Two methods are used (as suggested from the work undertaken in **Chapter 2**) to isolate the genetic signatures of events of positive selection in single loci. I calculate the F_{ST} values across the autosomal genome according to the Weir & Cockerham method in *VCFTOOLS* (Weir and Cockerham 1984; Danecek et al. 2011) pairwise for all populations vs Yoruba, as well as for all African population pairs. I then filtered to retain only biallelic sites and remove indels, and removed sites with low coverage, mapping quality and excess heterozygosity (Bergström et al. 2020).

The *Relate* programme was also implemented across the autosomal genome (Speidel et al. 2019), which requires phased input data in the format *haps/sample.* I filtered according to the same criteria given above (as well as removing SNPS with more than 10% missing data (Danecek et al. 2011)), before phasing using *SHAPEIT*2 (Delaneau et al. 2013). During phasing, I used the advised parameters of 0.3Mb window size and 200 conditioning states (number of conditioning haplotypes used during the phasing process; (Delaneau et al. 2013)). I identified eight chromosomes (chr9, 12 or 21) with more than 10% of data missing, but these are randomly distributed amongst individuals and therefore remain in the dataset.

The phased input files were then prepared for tree reconstruction according to the pre-processing steps in the *Relate* pipeline (Speidel et al. 2019). This includes flipping haplotypes according to an ancestral state (as taken from *ensemble* (Yates et al. 2020)), generating additional SNP annotations (the alleles upstream and downstream, as well as the number of carriers of the derived allele in each population, which are necessary for later estimates of population size using *Relate* (Speidel et al. 2019)) and adjusting

distances between SNPs (according to a genomic mask from (Bergström et al. 2020)). Following this, *Relate* was then used to reconstruct trees along the genome using the sample of 913 individuals (see **Section 3.3.2**). The effective population sizes throughout time, branch lengths and mutation rate were then simultaneously estimated to re-infer a tree for each locus. Finally, the programme was used to calculate the probability of variants at each locus rising to its observed frequency today, as given as a $-log_{10}pvalue$.

The *Relate* programme was also used to calculate the probabilities of positive selection acting on alleles on the X chromosome, with the following edits to the previously outlined method. I used the "phasing chromosome X" pipeline in $SHAPEIT2$ (which requires sex data) and remove one individual who has 75% of SNPs missing on their X chromosome (HGDP01208; Oroqen population). *Relate* is then ran as previously described but using the haploid input data files, treating each female as two haploid samples and each male as one haploid sample.

3.3.4. Isolating Monogenic Signatures of Positive Selection

The F_{ST} and *Relate* probabilities for each MA-gene were extracted, where the former is given as a value between 0 and 1 (where 0 indicates no genetic differentiation and 1 indicates complete differentiation) and the latter is given as a $-log_{10}pvalue$ (where -1.30103 is equivalent to a $pvalue$ of 0.05, interpreted here as, given the variant's inferred trajectory, the probability of the variant acting under neutrality as 0.05). F_{ST} and *Relate* probabilities are extracted for each MA-gene region, as well as for the 10kbp regions up- and downstream in order to capture additional signatures of positive selection outside the gene region but that may still be related to its function (*i.e.,* as the case with variants surrounding *LCT* conferring lactase persistence, albeit on an unusually long haplotype (Tishkoff et al. 2007)).

I use the empirical genome-wide background, built from all SNPs along the genome, for each population (or population pair in the case of F_{ST}) to identify SNPs that fall in the tails of the F_{ST} and *Relate* empirical distribution. Here, I assign SNPs with selection values in 0.1% tail as those with evidence of selection. When considering signatures of positive selection across an entire gene set, I also include SNPs with selection values in the 5% tail, as signatures of positive selection are expected to be weaker under polygenic adaptation. Whilst the $-log_{10}pvalue$ of *Relate* can be transformed and used explicitly as a $pvalue$, I choose to use the tail of the empirical distribution to identify candidate targets of positive selection since I have shown that this increases accuracy when using sample sizes under 50 (see **Chapter 2**).

The signatures of positive selection identified by these two methods are related but subtly different. The tail of the empirical F_{ST} distribution contains sites that are the most highly differentiated between populations (and hence can be expected to be enriched with targets of positive selection, since such differentiation is unlikely, although not impossible, under neutrality). A key subtly of F_{ST} ,therefore, is that it can only identify signatures of positive selection that have arisen following the split of the two populations used in each pairwise calculation. *Relate*, however, identifies sites that have risen to an unusual frequency, given their age and the number of lineages present when they first arose, over the entire inferred history of the locus in a given population (in reality, this is up to the time of the common ancestor of all populations used in the genealogical inference). When using the empirical distribution to identify outliers as

done so here, this is specifically identifying SNPs which have an unusually fast spread compared to all other SNPs within this population's inferred history (and hence can also be expected to be enriched for targets of positive selection). Therefore, the combination of these statistics allows, in theory, the identification of adaptation that has occurred differentially between populations and within the specific inferred history of an individual population.

3.3.5. Isolating Polygenic Signatures of Positive Selection

I assess if the entire MA-gene set is significantly enriched with signatures of positive selection, as identified by either F_{ST} or $Relate$. I use a $chi\ squared$ test to compare the number of SNPs at the 5% significance level to the neutral expectation (5% of all SNPs in the gene set). I repeat this for SNPs at the 1% significance level (where the neutral expectation is now 1% of all SNPs in the gene set). Finally, to more explicitly test for an excess of signatures of positive selection in each functionally-related set, I repeat this individually for each MA-gene set separately, testing for an enrichment of SNPs at the 5% significance level (for signatures of positive selection identified by either F_{ST} or $Relate$).

The gene-set enrichment method $SUMSTAT$ (Daub et al. 2013) is then applied to investigate the signatures of polygenic adaptation on individual MA-gene sets. Here, I extract the most extreme F_{ST} and $Relate$ probabilities for each MA-gene, and sum these across micronutrient gene sets to generate a summed MA-gene set value for each statistic. The summed MA-gene set values are then compared to the background set summed values generated from 1,000 neutral gene sets. The neutral gene sets are built from a random combination of the pMA-genes (see **Section 3.3.1.1**) corresponding to each MA-gene within the test MA-gene set. A $Python$ script is then used to identify MA-gene set summed values that fall in the 5% tail of this background distribution, as generated from these neutral set values (see **Supplementary Note S3.1**).

3.4. Results

3.4.1. Patterns of Adaptation in Micronutrient-Associated Genes

I begin by exploring the signatures of local positive selection across the entire micronutrient gene set. Since both monogenic and polygenic signatures of selection are of interest, I extract the SNPs within the 5% tail of the empirical distribution of either $Relate$ or F_{ST} for each MA-gene and for each population. As an additional precautionary step, I only consider MA-genes with more than five SNPs within the tail of the respective empirical distribution. I then identify the SNP with the strongest signature of positive selection in each MA-gene, which is considered the strongest candidate target SNP. I observe SNPs with these signatures of positive selection across all micronutrient categories and across populations of all major global geographic areas (see $Relate$ results in **Fig. 3.2A,** F_{ST} results in **Fig. 3.3A**). Notably, many MA-genes contain SNPs which fall in the extreme 0.1% tail of the empirical distribution of either $Relate$ or F_{ST}, the threshold for individual genes showcasing evidence for positive selection.

Prior to exploring these individual signatures of positive selection, and in recognition that not each SNP in the tails of the empirical background distribution is necessarily a true target of positive selection, I first investigate if there is an excess of signatures of

positive selection across the entire MA-gene set. Hence, I ask if there are more SNPs than expected at the 0.1% tail (*i.e.,* showing significant evidence of positive selection) and the 5% tail (*i.e.,* showing weak evidence of positive selection, as expected for example for polygenic adaptation within a gene set only) within each population.

Compared to neutral expectations, there is a significant excess of SNPs within the 0.1% tail of the empirical distribution of both $Relate$ or F_{ST} in many of the populations (**Fig. 3.4**). This excess is observed in more populations for F_{ST} than $Relate$. Moreover, a majority of populations also show a significant excess of SNPs within the 5% tail of the F_{ST} empirical distribution, but no populations show a significant excess of SNPs within the 5% tail of the $Relate$ empirical distribution. Still, this does not exclude the presence of strong signatures of positive selection across many genes in individual MA-gene sets (addressed in **Section 3.4.3**). Despite the limitations of this simple approach, which fails to account for the genomic structure of the SNPs (see **Section 3.4.3**) this analysis suggests higher than expected differentiation of MA-genes among populations, in line with expectations of positive selection within this gene set.

I now use the signatures of positive selection summarised in **Fig. 3.2** and **Fig. 3.3** to address preliminary questions. I first ask if the signatures of positive selection identified on MA-genes appear to be randomly distributed amongst micronutrients and populations, or if they cluster within certain micronutrient gene sets and/or certain populations. If signatures of positive selection cluster within a group of biologically related genes, *i.e.,* a MA-gene set, this can suggest adaptation of the corresponding micronutrient-associated function. In addition, if signatures of positive selection within a MA-gene set cluster in certain populations, this can indicate which populations may have undergone genetic adaptation. As an extension of this, I also ask if the geographical distribution of signatures of positive selection indicate whether putative adaptation to micronutrients is strictly local (*i.e.,* on the level of individual populations or continents) or more global (spread across multiple continents).

I first ask these questions with respect to the signatures of positive selection as identified by $Relate$. Here, many signatures within the same MA-gene set are observed very locally (*e.g.,* phosphorus-associated genes in the American Pima population) whilst others cluster across continental regions (*e.g.,* selenium-associated genes in East Asia). Other MA-genes show strikingly widespread geographic signatures of selection (*e.g.,* those identified in zinc-associated genes in non-African populations). Therefore, modern humans may have a history of both geographically global and local micronutrient-associated adaptations. Finally, the number of MA-genes exhibiting signatures of positive selection within each MA-gene set is highly variable, suggesting that the degree of polygenicity of micronutrient-associated adaptation likely also varies amongst micronutrients (polygenicity addressed in **Section 3.4.3**).

Before I ask the same questions with respect to the signatures of positive selection identified by F_{ST}, the signatures identified by F_{ST} must be considered to differ from the above signatures of positive selection identified by $Relate$. Whilst $Relate$ identifies SNPs with a trajectory improbable under neutrality, here F_{ST} identifies SNPs which are most highly differentiated in each individual population to the Yoruba population. Hence, for most pairwise combinations in this study, F_{ST} is used to identify SNPs with unusual differentiation between African (Yoruba) and non-African populations. Therefore, F_{ST} may capture signatures of positive selection localised to an individual

population, as well as signatures of positive selection that reflect adaptation in an ancestral non-African population or the Yoruba population.

Many signatures of positive selection identified by F_{ST} within the same MA-gene set are also observed both very locally (*e.g.,* iron-associated genes in the Oceanian Bougainville and America Pima populations) or at the continental (*e.g.,* calcium-associated genes in Europe and Central-South Asia) level. The most striking difference between the F_{ST} and $Relate$ signatures of positive selection however is that, in general, the F_{ST} signatures are shared over more populations, particularly non-African populations, compared to those of $Relate$ (see **Fig. 3.3A**). This is especially observed in some selenium, magnesium, zinc and phosphorus-associated genes, and indicative of potential associated selection events swiftly preceding, overlapping with or following the Out of Africa migration.

Finally, to consider the allele frequency differentiation between African populations that may indicate local adaptation events within Africa (and remove the limitation of only identifying extreme differentiation from Yoruba), the F_{ST} analysis is expanded to consider all cross-African population pairs (see **Fig. 3.3B**). Once more, signatures of positive selection are concentrated in certain micronutrient gene sets in individual populations (*e.g.,* calcium-associated genes in the African Biaka population), but do not show same widespread geographic distribution of signatures of positive selection as observed in **Fig 3.3A**. This is as expected, since these African populations do not have the same degree of shared history compared to non-Africans.

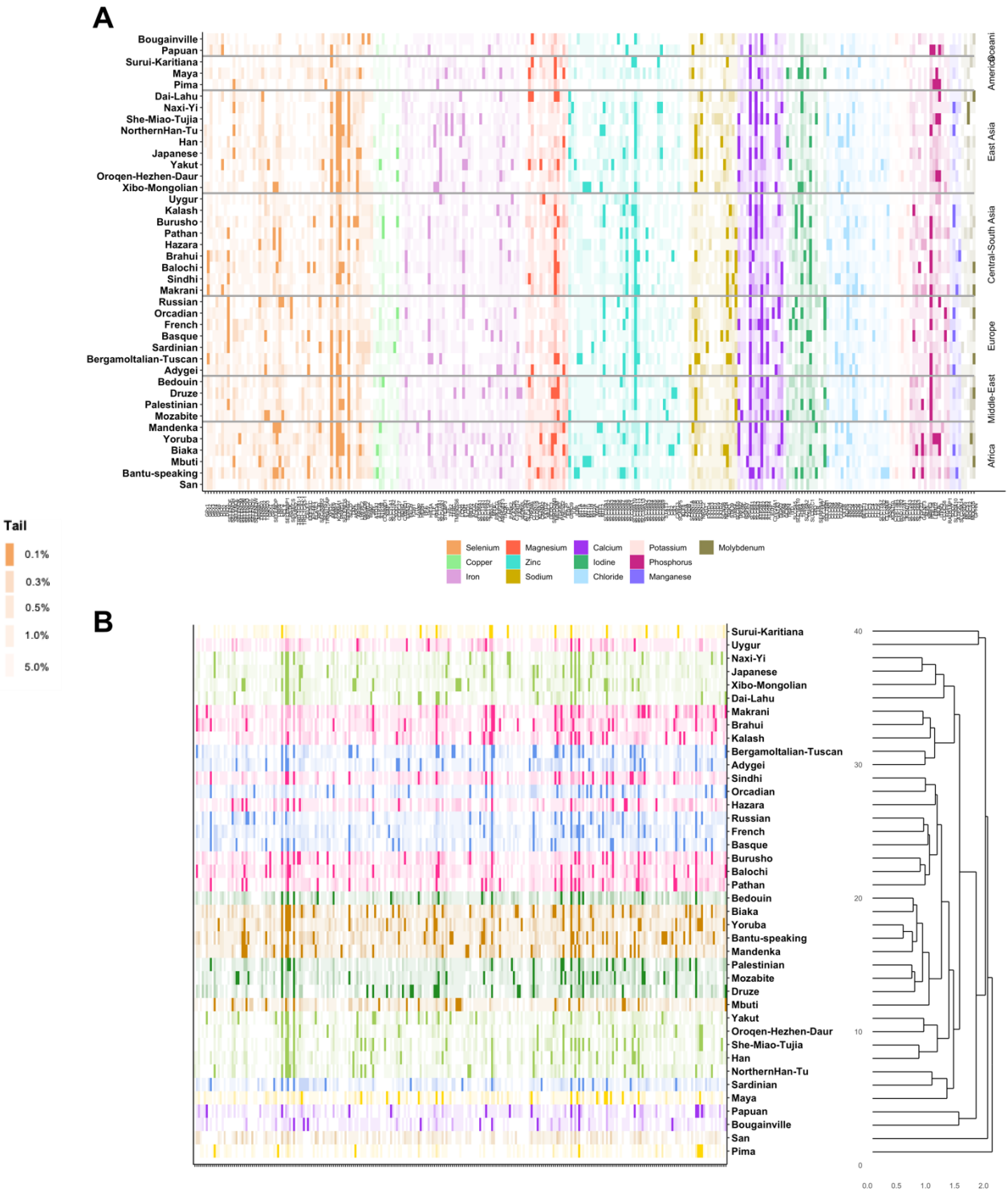


Figure 3.2: Relate signatures of positive selection over populations. *A) Strength of Relate signatures of positive selection across all MA-genes (x-axis, coloured by micronutrient) and all populations (y-axis, grouped by metapopulation). The darkness of the blocks (see left legend) reflects the strength of the signature (5%, 1%, 0.5%, 0.3%, 0.1% tails of the empirical distribution shown) with the darkest blocks indicating SNPs at the 0.1% tail. B) Strength of Relate signatures of positive selection across all MA-genes grouped according to the population clustering (right dendrogram) as calculated from the significance of the most extreme Relate probabilities over all MA-genes.*

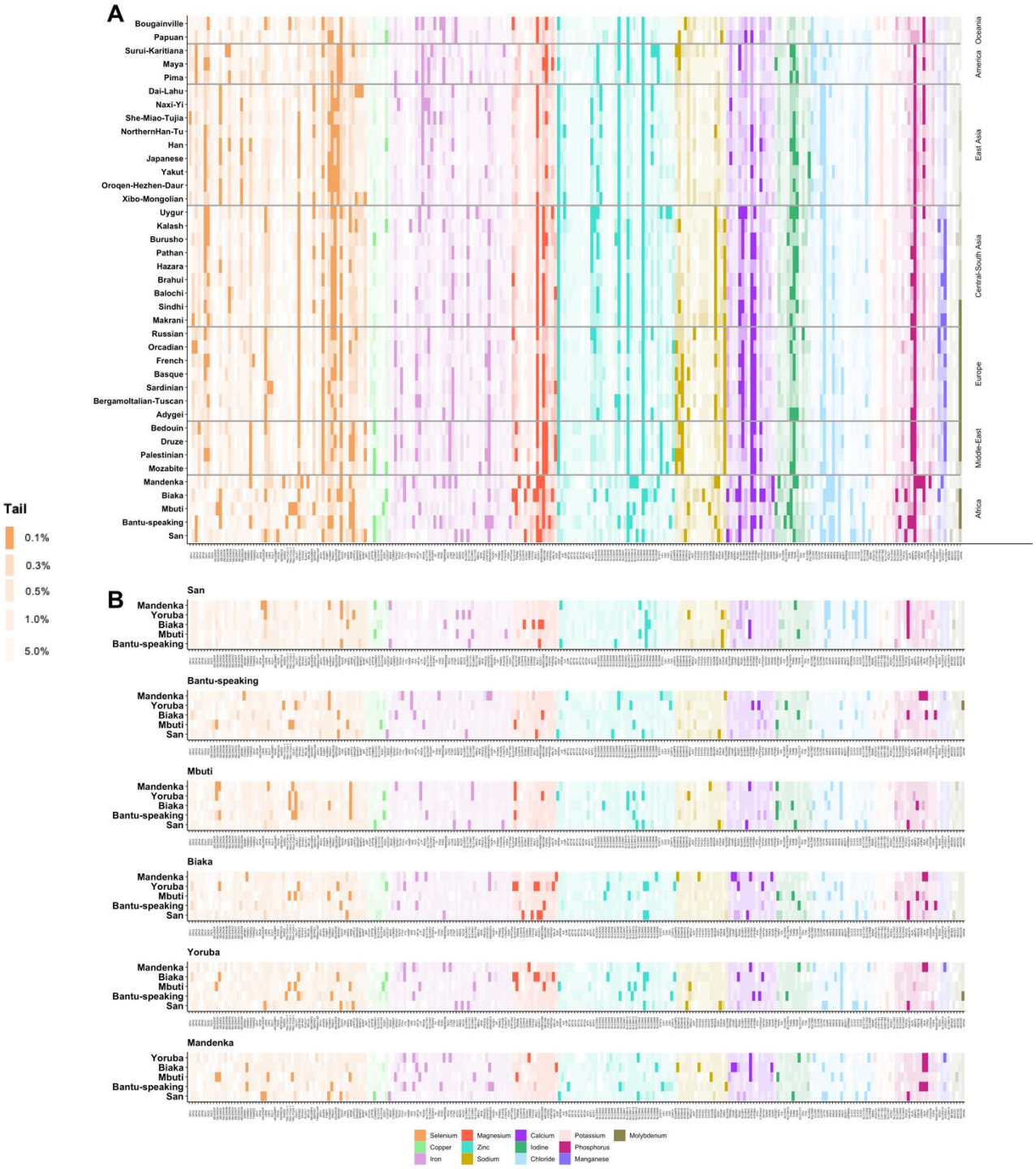


Figure 3.3: F_{ST} Signatures of positive election over populations *A) Strength of F_{ST} selection signatures of positive selection across all MA-genes (x-axis, coloured by dominant micronutrient association) and all F_{ST} pairwise comparisons with Yoruba (y-axis, grouped by metapopulation). B) Strength of F_{ST} selection signatures across all MA-genes (x-axis, coloured by dominant micronutrient association) and all F_{ST} pairwise comparisons amongst African populations (pairwise comparisons for each panel are those between the title population and those listed on the y-axis). The darkness of the blocks (see left legend) reflects the strength of the signature (5%, 1%, 0.5%, 0.3%, 0.1% tails of the empirical distribution shown) with the darkest blocks indicating SNPs at the 0.1% tail.*

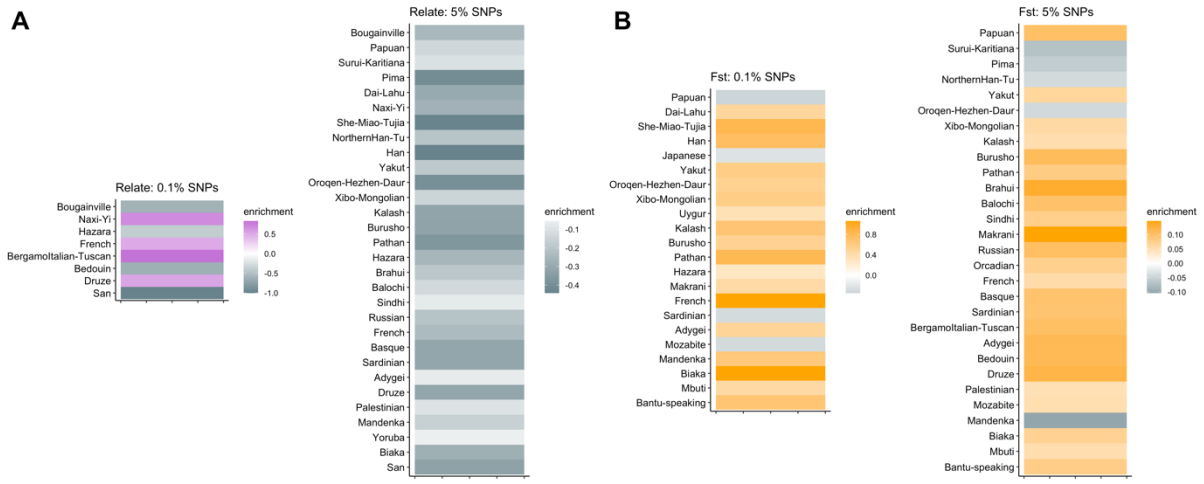


***Figure 3.4*: Populations showing a significant excess or deficit of SNPs of MA-genes.** *Significant excess or deficit as identified within the 0.1% and 5% tails of the empirical distribution of A) Relate and B) F_{ST} (populations with pvalue < 0.05 as calculated by a chi − squared test comparing the number of SNPs observed in the respective tail to the number of expected SNPs: either 0.1% or 5% of the total SNPs). Grey shows a significant deficit (less SNPs in the tail than expected), purple and orange show a significant excess (more SNPs in the tail than expected, for Relate and F_{ST} respectively).*

3.4.2. Adaptation Across Locality

I now investigate if the signatures of positive selection reflect the geography of the populations. To do so, I only use the signatures of positive selection identified by *Relate*, to avoid the loss of fine-scale local adaptation possible with the use of F_{ST} (which here most explicitly captures differentiation from the African Yoruba population).

For each population, each MA-gene is represented by the *pvalue* of the SNP with the strongest evidence for selection. I then cluster all populations according to the *pvalues* across all MA-genes using the *hclust* package of *R*; see **Fig. 3.2B**. From this, I can explore if the signatures over all MA-genes groups populations by geography and, by extension, which populations exhibit genetic signatures that are unusual compared to its geographically (and genetically) closest populations.

In general, populations group with geographically proximal populations; groups are formed from European and Central-South Asian populations; African and Middle-Eastern populations; northern East Asian populations and southern East Asian populations. Hence, this grouping mostly reflects shared ancestry. Interestingly, the American populations fail to cluster together (**Fig. 3.2B**). These populations are geographically further apart when compared to other continental groups (with the three American populations occupying land in Northern, Central and Southern America see **Fig. 3.1**) and it is thus possible that the different environmental conditions experienced by each population have resulted in the differential genetic signatures here,

rather than shared ancestry or demographic history. Still, this remains only a speculation and an interesting outlier to the broad pattern of population clustering.

I now focus on the SNPs with evidence of positive selection (within the 0.1% tail of the empirical distribution of *Relate*) and test if they show the same extent of geographical structure, and how this compares between MA-gene sets. For each MA-gene set, I calculate the mean number of MA-genes showing significant evidence of positive selection in each metapopulation, calculate the standard deviation (to represent the variance within each metapopulation) and normalise the mean over each micronutrient category (to compare between micronutrients). From this, I am able to preliminarily investigate if the evidence of positive selection associated with each micronutrient appears to be shared across continental space, or if they may instead be localised within individual populations[3].

The number of MA-genes showing evidence of positive selection are, on average, highly variable within metapopulations (large standard deviations; see **Table 3.2**). Regionally and genetically-close populations do not show similar numbers of MA-genes with signatures of positive selection within each MA-gene set, indicating that the strongest signatures of selection are at the local scale. A potential exception to this is the number of selenium-associated genes with evidence of positive selection, which is the highest (with the lowest variance) in the African metapopulation, followed by the East Asian metapopulation (with a comparatively low variance compared to other populations). This is in accordance with selenium-associated selective pressures shared over many populations in these continents (Hurst et al. 2013; White et al. 2015).

[3] We advise that because of the low number of populations within certain metapopulation groups, the results of this analysis should be focused on the European, Central-South Asian and East Asian metapopulations.

***Table 3.2**: **Summary statistics for each MA-gene set by metapopulation.** Mean, normalised mean, standard deviation ("SD") and maximum ("Max.") number of MA-genes with Relate signatures of positive selection (within the 0.1% tail of the empirical distribution) for each micronutrient gene set across each metapopulation. San population was removed from the African metapopulation due to its sample size (n=6). Se=selenium, Cu=copper, Fe=iron, Mg=magnesium, Zn=zinc, Na=sodium, Ca=calcium, I=iodine, Cl=chloride, K=potassium, P=phosphorus, Mn=manganese, Mb=molybdenum.*

	Mean	Normalised mean	SD	Max.
Se				
Africa	6.60	0.35	0.89	8
Middle-East	3.75	-0.01	1.50	5
Europe	4.29	0.06	2.14	7
Central-South Asia	3.56	-0.03	2.24	6
East Asia	4.78	0.12	1.39	7
America	1.33	-0.36	0.58	2
Oceania	2.5	-0.17	0.71	3

	Mean	Normalised mean	SD	Max.
Cu				
Africa	0.6	0.04	0.55	1
Middle-East	0.5	0.09	0.58	1
Europe	0.29	-0.02	0.49	1
Central-South Asia	0.33	0	0.71	2
East Asia	0.33	0	0.5	1
America	0	0.17	0	0
Oceania	0	0.17	0	0

	Mean	Normalised mean	SD	Max.
Fe				
Africa	2.6	0.11	1.52	5
Middle-East	2.25	0.05	2.63	6
Europe	2.43	0.08	1.13	4
Central-South Asia	1.56	-0.07	0.73	3
East Asia	2.56	0.1	1.42	4
America	1.33	-0.11	1.53	3
Oceania	1	-0.16	1.41	2

	Mean	Normalised mean	SD	Max.
Mg				
Africa	2	0.145	1	3
Middle-East	1.75	0.08	1.26	3
Europe	1	-0.11	1.55	3
Central-South Asia	1.44	0.01	0.88	3
East Asia	1.22	-0.05	1.48	4
America	2	0.15	1.73	3
Oceania	0.5	-0.23	0.71	1

	Mean	Normalised mean	SD	Max.
Zn				
Africa	4.2	0.26	0.84	5
Middle-East	2.5	-0.03	0.58	3
Europe	3.29	0.10	1.11	5
Central-South Asia	2.78	0.02	1.78	6
East Asia	2.89	0.04	2.02	6
America	1	-0.28	1.73	3
Oceania	2	-0.11	0	2

	Mean	Normalised mean	SD	Max.
Na				
Africa	1.4	-0.05	1.14	3
Middle-East	2.25	0.19	0.96	3
Europe	2.72	0.28	1.25	4
Central-South Asia	1.89	0.08	1.70	4
East Asia	1.22	-0.09	0.67	2
America	0.67	-0.23	0.58	1
Oceania	1	-0.15	0	1

	Mean	Normalised mean	SD	Max.		Mean	Normalised mean	SD	Max.		Mean	Normalised mean	SD	Max.
Ca					**I**					**Cl**				
Africa	4	0.24	1.22	5	Africa	1.6	0.04	1.14	3	Africa	2.8	-0.21	1.30	4
Middle-East	3	0.07	0.82	4	Middle-East	1.25	-0.032	0.96	2	Middle-East	2.25	0.10	1.5	4
Europe	3.14	0.10	1.35	5	Europe	2	0.12	1.15	3	Europe	2.29	0.10	0.76	3
Central-South Asia	2.22	-0.06	1.20	5	Central-South Asia	1.44	0.01	0.88	3	Central-South Asia	2.22	0.09	1.64	5
East Asia	2.78	0.04	1.86	6	East Asia	1.11	-0.06	0.79	2	East Asia	1.67	-0.02	1.22	3
America	1.33	-0.21	0.58	2	America	2	0.12	2.65	5	America	0.67	-0.22	0.58	1
Oceania	1.5	-0.18	0.71	2	Oceania	0.5	-0.18	0.71	1	Oceania	0.5	-0.25	0.71	1

	Mean	Normalised mean	SD	Max.		Mean	Normalised mean	SD	Max.		Mean	Normalised mean	SD	Max.
K					**P**					**Mn**				
Africa	1	0	0.71	2	Africa	1.80	0.14	1.79	4	Africa	0.4	0.05	0.55	1
Middle-East	1.5	0.25	1	2	Middle-East	1.5	0.06	0.58	2	Middle-East	0	-0.15	0	0
Europe	1.43	0.22	0.79	2	Europe	1.14	-0.03	0.69	2	Europe	0.57	0.14	0.53	1
Central-South Asia	0.89	-0.06	0.78	2	Central-South Asia	0.78	-0.12	0.67	2	Central-South Asia	0.44	0.07	0.53	1
East Asia	0.56	-0.22	0.53	1	East Asia	0.67	-0.15	0.87	2	East Asia	0.33	0.02	0.71	2
America	0.67	-0.17	0.58	1	America	1.33	0.02	1.53	3	America	0.33	0.02	0.58	1
Oceania	1	0	0	1	Oceania	1.5	0.06	2.12	3	Oceania	0	-0.15	0	0

	Mean	Normalised mean	SD	Max.
Mb				
Africa	0.4	0.19	0.55	1
Middle-East	0.25	0.04	0.5	1
Europe	0.29	0.08	0.49	1
Central-South Asia	0.22	0.01	0.44	1
East Asia	0.33	0.12	0.5	1
America	0	-0.21	0	0
Oceania	0	-0.21	0	0

3.4.3. Assessing the Polygenicity of Selection

3.4.3.1. Adaptation over Individual MA-Gene Sets

There is some, but limited, evidence of positive selection across the entire MA-gene set, as well as multiple MA-genes within the same individual MA-gene sets exhibiting evidence of positive selection (**Section 3.4.1**). It is hence possible that individual MA-gene sets have undergone oligogenic or polygenic adaptation in response to micronutrient-associated pressures, which has resulted in the limited excess of significant SNPs over all MA-genes. To explore this possibility, I test each MA-gene set individually for an excess of SNPs with signatures of positive selection. I evaluate if the number of SNPs in the 5% tail of the empirical distribution of either F_{ST} or $Relate$ (using the less stringent tail since weaker signatures of positive selection can be expected to accompany polygenic adaptation) is higher than the expected 5% of total SNPs for each micronutrient gene set (**Fig. 3.5**).

An excess of significant SNPs within multiple micronutrient gene sets is observed in many populations (and suggesting that power is gained when examining each MA-gene set individually). This includes an extreme excess of significant SNPs according to both F_{ST} and $Relate$ in the molybdenum gene set, in European, Central-South Asian, Middle-Eastern and some African populations (**Fig. 3.5B**). This appears to be driven by the high number of significant SNPs of the *GPHN* and *MOCS2* genes, two of the only five genes in this gene-set. Hence, this excess of significant SNPs is likely a result of linkage disequilibrium and I caution against interpreting this as a signal of polygenic adaptation. Since the excess of significant SNPs in the molybdenum gene set is so extreme when compared to other gene sets, the molybdenum gene set is removed from **Fig. 3.5A** before evaluating the other excesses of significant SNPs amongst populations.

An excess of significant SNPs identified by $Relate$ is observed in all MA-gene sets, bar iron, for at least one population (despite only four populations showing an excess of significant SNPs identified by $Relate$ when considering all MA-genes, see **Section 3.4.1, Fig. 3.4**). This includes an excess of significant SNPs in six populations for each of copper, zinc and calcium-associated genes and five populations for iodine-associated genes; an excess of significant SNPs in the selenium-associated genes in two East Asian populations (Yakut and Xibo-Mongolian); and high significant excesses of significant SNPs in the MA-gene sets of individual populations (*e.g.,* 79% more significant SNPs than expected in iodine-associated genes in the American Maya, and 104% more significant SNPs than expected in manganese-associated genes in the American Surui-Karitiana). Whilst the role of linkage disequilibrium in driving these cases of excess significant SNPs cannot be discounted, they remain interesting signatures of which warrant further exploration.

Some of these cases of excess significant SNPs are also captured by the analogous F_{ST} analysis. For example, there is also an excess of significant SNPs in many populations for the copper, zinc, calcium and iodine-associated genes sets, as well as an excess number of significant SNPs in the iodine-associated genes of the Maya (**Fig 3.5**). Moreover, the populations with an excess of significant SNPs as identified by F_{ST} in selenium-associated genes also includes the East Asian Yakut and Xibo-Mongolian populations (and extends to all other populations of East Asia). When an excess of significant SNPs is observed according to both the $Relate$ and F_{ST} signatures of positive selection, these

populations are stronger candidates for undergoing polygenic adaptation in response to the levels of their respective micronutrients.

Finally, recapitulating the observations from **Fig. 3.4**, the excess of significant SNPs according to the F_{ST} signatures of positive selection are more widespread than those identified according to $Relate$. This is particularly the case for the zinc, iodine and potassium gene sets, which show the strongest evidence for unusual differentiation from the African Yoruba population at the gene set level. Such differentiation between the African Yoruba and multiple non-African populations may be the result of selection on these gene sets in the Yoruba population, or more ancient selection in an ancestral non-African population.

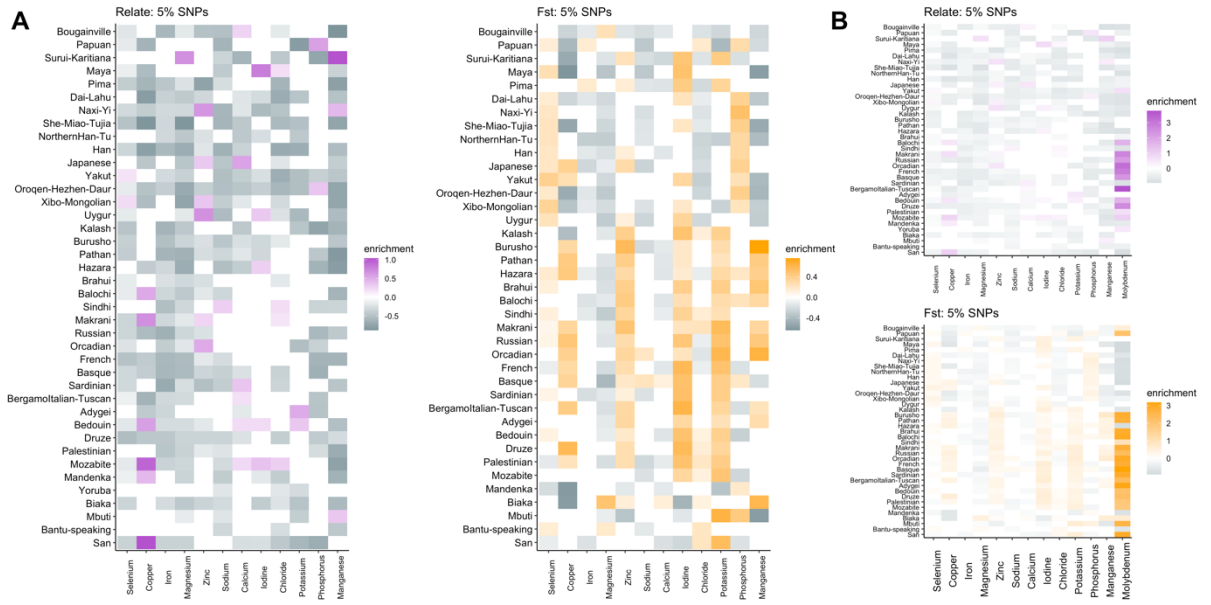


Figure 3.5: Populations showing a significant excess or deficit of SNPs of MA-genes for each MA-gene set.*The excess of significant SNPs, as identified within the 5% tails of the empirical distribution of* $Relate$ *and* F_{ST}*, for each population and micronutrient gene set. Significance calculated by a* $chi-squared\ test$ *(comparing the number of SNPs observed in the 5% tail to the expected 5% of total SNPs); grey shows a significant deficit (less SNPs in the tail than expected), purple and orange show a significant excess (more SNPs in the tail than expected, for* $Relate$ *and* F_{ST} *respectively). A) gives the results for all gene sets excluding molybdenum; B) includes the molybdenum gene set.*

3.4.3.2. Polygenic Adaptation over Individual MA-Gene Sets

For each gene set, the analysis thus far only tests if there is an excess of significant SNPs over all genes within a micronutrient gene set, and does not explicitly test if there are signatures of positive selection over a significant proportion of all genes within a gene set. The latter case would be the expectation under a classic model of polygenic adaptation. To investigate if the signatures of positive selection across micronutrient

gene sets agree with such a model, and if there is further evidence of which is necessary for a more conclusive claim of polygenic adaptation, I carry out the gene set method $SUMSTAT$ (Daub et al. 2013).

$SUMSTAT$ is first applied using the full micronutrient gene sets (*i.e.,* some overlap between gene sets; see **Tables S3.1**), summing the most significant $pvalues$ over all MA-genes within a gene set, and comparing this to the neutral background (see **Section 3.3.5**). For most MA-gene sets, multiple populations have summed MA-gene set values in the 5% extreme tail of the neutral distribution of $SUMSTAT$ values integrating either F_{ST} or $Relate$ signatures of positive selection (**Tables S3.7-8**). The strongest evidence of polygenic adaptation is observed in the phosphorus gene set of the Pima population ($pvalue$: 0.000013 , $SUMSTAT$ integrating $Relate$ signatures of positive selection), which was also suggested from **Fig. 3.2**.

Other populations with the strongest evidence of polygenic adaptation in response to micronutrient-associated pressures are those with $SUMSTAT$ values significant for the same micronutrient gene set when integrating either F_{ST} or $Relate$ signatures of positive selection. This is the case for the selenium, sodium and potassium gene sets (for one, seven and nine populations respectively; see **Table S3.9**). Here, the selenium gene set is significant when integrating either statistic in the Xibo-Mongolian, further suggesting a degree of polygenic adaptation in response to selenium-associated pressures in this population (as suggested in other East Asian populations from previous studies (Hurst et al. 2013; White et al. 2015).

However, the significance of $SUMSTAT$ gene set values are all below the multiple testing threshold ($\frac{0.05}{40\times13}$, or $p \leq 9.62e^{-5}$). Moreover, when repeating the $SUMSTAT$ method on the cut-down micronutrient gene sets (*i.e.,* those with no overlap, in order to avoid false positives in one gene set driven by signatures associated with another micronutrient), virtually all $SUMSTAT$ significant signatures disappear (excluding the selenium gene set in the Xibo-Mongolian; see **Table S3.10-11**). This indicates that the limited signatures of polygenic adaptation inferred from $SUMSTAT$ are strongly influenced by a small number of genes, those that are functionally associated with multiple micronutrients. Therefore, at the given power of these methods, there is insufficient evidence for a classic model of polygenic adaptation amongst micronutrient gene sets, where selection acts over the entirety, of significant proportion of a gene set.

Still, the polygenic analysis thus far, and the presence of signatures of positive selection across multiple MA-genes for virtually all micronutrient sets, does suggest that micronutrient-associated adaptation may be frequently mediated by more than one gene. Hence, I suggest the presence of polygenic adaptation on the scale of fewer genes, otherwise referred to as oligogenic adaptation. To explore this, and to further understand if the evidence of positive selection on MA-genes ever stretches across the majority of any MA-gene set, I calculate the proportions of MA-genes showing signatures of positive selection in each MA-gene set according to either $Relate$ or F_{ST} in each population.

Indeed, the proportions of MA-genes showing evidence of positive selection is never above 50% for any micronutrient gene set, using either the $Relate$ or F_{ST} evidence for selection (where few are above 20% when considering the $Relate$ signatures of positive selection; see **Fig. 3.6A**). Notably, this includes the case of the selenium gene set in the Xibo-Mongolian population (although, this is somewhat biased by the large selenium gene set size, n=59), and could further suggest that the previously calculated

SUMSTAT pvalues are a result of strong evidence of selection amongst only a few genes. From this, and the other results presented in this section, I conclude that the signatures of positive selection are not shared over the majority of any MA-gene set and, by extension, adaptation is likely mediated by only a small number of genes and unlikely to be classically polygenic in nature.

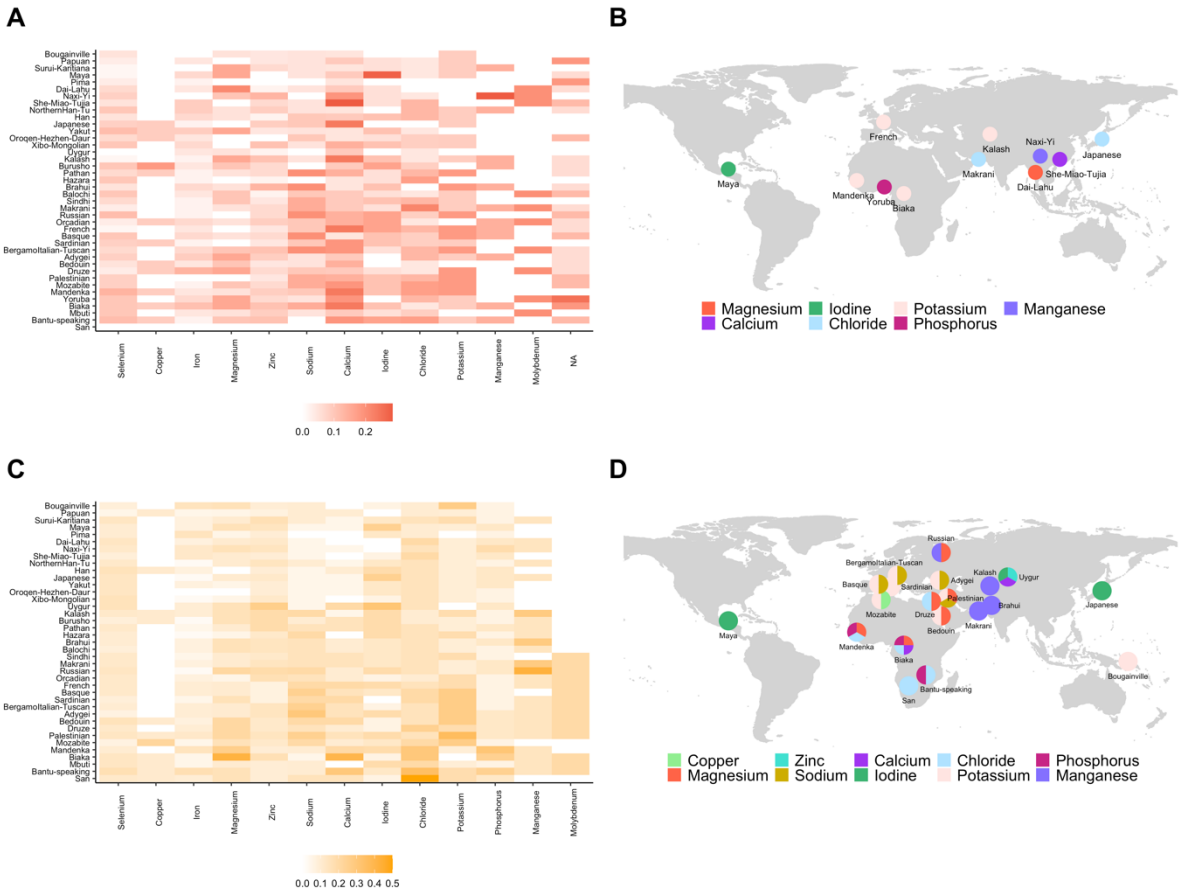


Fig. 3.6: Proportion of MA-gene sets with signatures of positive selection. *A) and B) show the proportion of each micronutrient gene set that have signatures in the 0.1% tail for Relate and F_{ST} selection values, respectively. Keys below the panels show the proportion of genes within a gene set with such signatures. F_{ST} gene sets do not include genes on the X chromosome. C) and D) show the populations with MA-gene sets with more than 20% of their genes exhibiting signatures of positive selection, with these MA-gene sets represented by the colours (given in the key below) in the population's corresponding circle.*

3.4.4. Candidate Populations for Micronutrient Adaptation

I propose the populations that are most likely to have undergone micronutrient-associated adaptation via two main avenues. I first identify the populations that show the highest proportion of MA-genes with signatures of positive selection in each micronutrient category, hence assuming a degree of oligogenic adaptation. Of the 40 populations, 25 have at least one micronutrient gene set with $>$ 20% of genes showing signatures of positive selection according to either the *Relate* or F_{ST} empirical distributions (see **Fig. 3.6C, D**). This includes the iodine-associated genes in the

American Maya (27.8% and 25% of genes showing signatures of positive selection according to $Relate$ or F_{ST}, respectively), calcium-associated genes in Central-south Asian populations and the East Asian She-Miao-Tujia (the latter having 28.6% of genes identified with signatures of positive selection, the highest $Relate$ proportion), and chloride in African populations, particularly the San (50%, the highest F_{ST} proportion).

For the second approach, I extend this by only now considering the MA-genes with the very strongest evidence of selection for each population, according to either $Relate$ or F_{ST} signatures of positive selection. I extract the MA-genes with the top five strongest signatures of selection (which may be more than five MA-genes when the signatures are of the same strength) and identify populations which show a clustering of these strongest signatures according to micronutrient category (see **Fig. 3.7, Tables S3.12-13**). Hence, I identify populations with repeatedly strong evidence for adaptation associated with the same micronutrient.

Five of the six highest ranked MA-genes in the Central-South Asian Hazara, according to the $Relate$ signatures of positive selection, are associated with selenium (with other populations showing high numbers of selenium-associated genes with this strong evidence of positive selection being the East Asian Oroqen-Hezhen-Daur and the African Mbuti and San populations). Other populations showing multiple top-ranking genes assigned to the same micronutrient include the American Pima (many phosphorus-associated genes), Middle-Eastern Bedouin and European Basque and French (all showing many iron-associated genes) and the African Mandenka and East-Asian Dai-Lahu (both showing many calcium-associated genes). I therefore present these populations as candidates for undergoing adaptation in response to these respective micronutrient levels.

The highest-ranking MA-genes according to F_{ST}, however, show a far more striking geographic pattern. Here, zinc-associated genes are amongst the top five ranks of MA-genes for the majority of Eurasian populations, particularly for Asian populations. This is suggestive of shared selection across these populations or, considering the signatures of positive selection identified by F_{ST} being those of differentiation between these populations and the African Yoruba population, possibly selection acting on zinc-associated genes following the Out of Africa migration. Selenium-associated genes also commonly rank as the MA-genes with the strongest evidence of selection according to F_{ST}, particularly in some East Asian and African populations, and may also indicate positive selection shared across continents.

In summary of the consideration of polygenic adaptation, I show that many populations in different global areas demonstrate evidence for mediating micronutrient-associated pressures via adaptations of multiple genes. This evidence stems from either 1) populations showing more evidence of selection across their gene set than expected under neutrality (as calculated by the excess of significant SNPs and $SUMSTAT$ analysis, see **Section 3.4.3**); 2) populations exhibiting signatures of positive selection across what is deemed an unusual number of MA-genes within a MA-gene set; 3) or the strongest evidence of positive selection within a population observed in multiple MA-genes within the same MA-gene set. Notably, the iodine-gene set of the American Maya fulfils all three of these criteria, and I suggest this population as the strongest candidate for undergoing micronutrient-associated adaptation. Ultimately, I propose that populations are more suitably described as undergoing oligogenic adaptation rather

than polygenic adaptation, and largely mediate micronutrient-associated selective pressures via only a small number of genes.

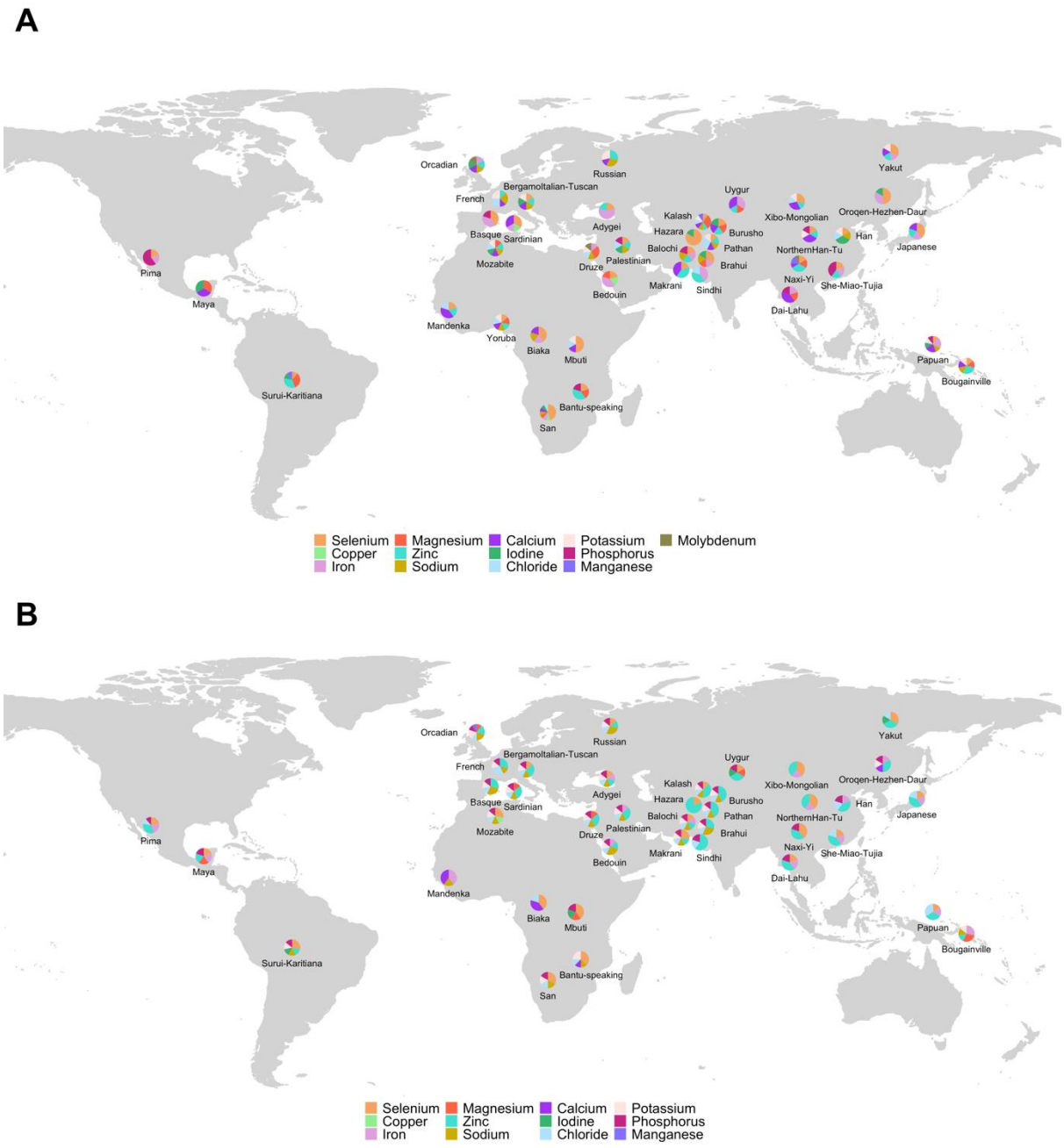


Fig. 3.7: ***The MA-genes with the top five strongest signatures of positive selection.*** *For each population, the micronutrient categories are given, coloured according to the key below, for the MA-genes with the top five strongest signatures of positive selection according to A) Relate and B)* F_{ST} *selection values.*

3.4.5. Candidate Target Genes for Positive Selection

Having identified the populations with the strongest evidence of positive selection (under the assumption of oligogenic adaptation) the MA-genes with the strongest evidence of positive selection are now considered. These candidate genes may make up the hypothesised small number of genes driving gene-set signatures (**see Sections 3.4.3-4**) or represent a monogenic signature of selection.

To identify the individual MA-genes most likely to have responded to micronutrient-associated selective pressures, I first extracted the MA-genes and their corresponding populations which have evidence of positive selection above the multiple-testing threshold ($\frac{0.05}{40\times269}$, or $p \leq 4.65e^{-6}$). 21 of the 40 populations show at least one MA-gene with signatures at this threshold according to either *Relate* or F_{ST}, distributed over 15 MA-genes associated with all micronutrients bar selenium, copper, manganese and molybdenum (**Table 3.3**). Of these 15 genes, *SLC12A1* (associated with potassium, sodium and chloride), *PDE7B* (associated with phosphorus) and *ATP2B2* (associated with calcium) show these strong signatures shared across populations, with the former two showing strong signatures of positive selection particularly in Middle-eastern and European/Central-south Asian populations. The absence of these strong signatures in other populations does not mean that selection in those populations should be discounted, only that the evidence for selection is weaker.

***Table 3.3*: Populations and their MA-genes with p-values below the multiple testing threshold of $\mathbf{4.65e^{-6}}$ (given in bold).** *Given alongside their associated micronutrient and accompanied by the p-value calculated by the other method to identify selection.*

Population	Gene	Micronutrient	*Relate* Significance	F_{ST} Significance
San	*GALNT3*	phosphorus	0.0849	**3.5e-6**
Mandenka	*ATP2B2*	calcium	0.000187	**7.75e-8**
Palestinian	*SLC12A1*	sodium, chloride, potassium	0.000137	**6.37e-7**
	THRB	iodine	**3.23e-6**	0.00159
Druze	*SLC12A1*	sodium, chloride, potassium	3.61e-05	**2.97e-7**
	PDE7B	phosphorus	0.000586	**8.16e-7**
Bedouin	*SLC12A1*	sodium, chloride, potassium	0.00571	**1.25e-6**
Adygei	*SLC12A1*	sodium, chloride, potassium	0.00481	**2.01e-6**
BergamoItalian-Tuscan	*SLC12A1*	sodium, chloride, potassium	0.0128	**1.58e-6**

	PDE7B	phosphorus	0.000678	**2.92e-6**
Sardinian	*ATP2B2*	calcium	**2.10e-7**	0.000969
	PDE7B	phosphorus	0.00366	**7.02e-7**
Basque	*SLC12A1*	sodium, chloride, potassium	0.000855	**2.00e-6**
	PDE7B	phosphorus	0.00522	**2.00e-6**
	HIF1A	iron	**2.43e-6**	0.000249
French	*SLC12A1*	sodium, chloride, potassium		7.65e-7
	SCNN1D	sodium, potassium	**1.87e-6**	0.000684
	PDE7B	phosphorus	0.000500	**4.28e-6**
Orcadian	*PDE7B*	phosphorus	0.00194	**2.24e-6**
Russian	*SLC12A1*	sodium, chloride, potassium	0.000519	**1.40e-6**
	SLC4A5	sodium	**3.83e-6**	0.000160
Makrani	*SLC39A11*	zinc	**1.40e-6**	0.0003304
	SLC39A4	zinc	0.0934	**3.95e-6**
Balochi	*SLC12A1*	sodium, chloride, potassium	0.0129	**1.76e-6**
Brahui	*SLC12A1*	sodium, chloride, potassium	0.00106	**1.07e-6**
	MECOM	magnesium	**1.26e-6**	0.000225
	PDE7B	phosphorus	0.000618	**1.53e-6**
Kalash	*SLC12A1*	sodium, chloride, potassium	0.0504	**2.56e-6**
Uygur	*FXYD2*	magnesium	**2.80e-6**	0.00923
Yakut	*FTMT*	iron	**3.37e-6**	0.000827
Han	*SLC30A9*	zinc	0.00228	**3.55e-6**
She-Miao-Tujia	*MLN*	phosphorus	**4.27e-6**	0.00460
Pima	*ATP2B2*	calcium	**9.06e-6**	0.00253

A MA-gene may still be considered a candidate for positive selection if consistently ranking amongst the MA-genes with the strongest evidence of selection over many populations, even if not reaching the multiple testing threshold within a single population. Hence, I also isolate the top-ranking MA-genes for each population and compare amongst all populations (**Table S3.12-13**). Many populations share the same MA-gene as that with the top-ranking evidence of positive selection, to the extent that only nine different MA-genes are represented as the top-ranking MA-gene across all non-African populations (according to the F_{ST} signatures of positive selection). This includes *SLC12A1* and *PDE7B*, but also the zinc-associated *SLC30A9* and *SLC39A4*, across

many and all Asian populations, respectively. The highest-ranking MA-genes according to the *Relate* selection values are more variable, but *PRKG1* (selenium-associated, with the strongest signature of selection over four East Asian populations), *ATP2B2*, *SLC8A3* and *SLC8A1* (calcium-associated), *SLC39A11* (zinc-associated) and *HIF1A* (iron-associated) are also shared as the highest-ranking MA-gene over multiple populations.

Hence, I present these genes as strong candidates for mediating micronutrient-associated adaptation shared across multiple populations. I particularly suggest *SLC12A1, PDE7B, SLC30A9, SLC39A4, ATP2B2, SLC3A11* and *HIF1A* as likely candidates of positive selection since, alongside the sharing of top-ranking signatures of positive selection amongst populations, they bypass the most stringent threshold in at least one population (see **Table 3.3**).

Finally, I consider the MA-genes which show signatures of positive selection shared across many populations but do not reach either the most stringent threshold or rank as exhibiting the strongest evidence within a population. MA-genes identified here, but not by the previous two criteria, may still represent candidates for monogenic adaptation, but I propose that their slightly weaker signatures of positive selection could more likely represent their role in mediating oligogenic adaptation shared across populations. In total, 49 MA-genes show *Relate* or F_{ST} signatures of positive selection in ten or more populations (**Fig. 3.8, Table S3.14**), which is the final set of candidate genes mediating adaptation shared across populations in response to micronutrient-associated pressures, either at the monogenic or oligogenic level.

Of these, *EEFSEC, PRKG1* (selenium-associated), *SLC39A11*, *SLC39A4, SLC30A9, GPR39, SLC39A11* (zinc-associated), *ATP2B2* (calcium-associated), *AQP6* (chloride-associated), *DCDC1* (magnesium-associated), *PDE7B* (phosphorus-associated), *TSHR* (iodine-associated) and *SLC12A1* (sodium, potassium and chloride-associated) share signatures of positive selection identified by F_{ST} in 20 or more non-African populations, thereby showing shared unusual differentiation to the African Yoruba population. There are limited signatures of positive selection identified by F_{ST} calculated between Yoruba and the remaining African populations for these same genes, hence these signatures may represent shared positive selection acting on a non-African common ancestor (rather than on positive selection acting on Yoruba). In support of a singular selection event, I also observe that it is the same SNP identified as having the strongest evidence of positive selection in over 10 populations in *SLC39A11* (rs11077654; in 10 populations according to *Relate*), *SLC39A4* (rs1871534; in 35 populations according to F_{ST}), *PDE7B* (rs7753890; in 35 populations according to F_{ST}) and *SLC12A1* (rs2413887; in 18 different populations according to F_{ST}). I therefore propose that *SLC39A11, SLC39A4, PDE7B* and *SLC12A1* are the strongest candidate genes for mediating micronutrient-associated adaptation surrounding or swiftly following the Out-of-Africa migration.

3.5. Discussion

Diet is highly variable across human populations, dictated by food availability, culture and soil geology (Xia et al. 2005; Tishkoff et al. 2007a; Minster et al. 2016; Hengl et al. 2017; Dhaliwal et al. 2019; De Groote et al. 2021). It has long been known that diet has played a role in human evolution (Osier et al. 2002; Han et al. 2007; Perry et al. 2007; Tishkoff et al. 2007a; Fumagalli et al. 2015; Schlebusch et al. 2015; White et al. 2015; Zhang et al. 2015a; Roca-Umbert et al. 2022), but the extent and dynamics of the selective impact of many dietary components remains unknown. Micronutrients are an essential

component of the human diet which, alongside their variability across global soils (Xia et al. 2005; Hengl et al. 2017; Dhaliwal et al. 2019; De Groote et al. 2021) makes them a strong candidate for driving local adaptation in modern humans.

Here, I present the first study to comprehensively investigate adaptation in response to the levels of 13 essential micronutrients across 40 modern human populations spanning every major area of the globe. Using the manually curated novel gene sets associated with each micronutrient, totalling 276 genes, I am able to evaluate the evidence of positive selection at the level of the entire MA-gene set, each individual MA-gene set and at the level of individual MA-genes. Hence, I am able to comprehensively evaluate the hypothesis that adaptation to micronutrients has driven genetic adaptation in modern human populations, either at the strictly local or at the more global scale.

3.5.1. Evidence for Micronutrient-Associated Adaptation

Firstly, signatures of natural selection are present across all micronutrient categories, and observed in all 40 of the global populations. There is a significant excess of signatures of positive selection across the entire MA-gene set in the majority of populations (**Section 3.4.1**), as well in individual MA-gene sets (**Section 3.4.3**). These excesses of signatures of positive selection, the co-occurrence of signatures of positive selection amongst many genes within functionally-related gene sets (**Section 3.4.3-4**), and the strong signatures of positive selection identified in individual MA-genes (**Section 3.4.5**) suggest the presence of, perhaps extensive, adaptation in modern humans associated with micronutrients.

Moreover, the geographical distribution of the signatures of positive selection are often supported by known soil deficiencies and toxicities across localities (Silvertooth et al. 2001; Vyshpolsky et al. 2008; Hurst et al. 2013; Ryan et al. 2013; White et al. 2015; Nell and van Huyssteen 2018; Hou et al. 2022). Previous studies have identified relationships between the signatures of positive selection and the environment of candidate populations (i.e., the soil levels of the relevant micronutrient), some of is recapitulated here (Hurst et al. 2013; White et al. 2015; Zhang et al. 2015a). I also identify additional novel cases where the signatures of positive selection within certain populations are supported by the known soil composition of their respective environment (see **Section 3.5.3-4**). Since the micronutrient composition of local soil affects the levels of micronutrients uptaken in the diet, this study provides preliminary link between this selective driver and signatures of micronutrient-associated adaptation.

3.5.2. Polygenicity of Micronutrient-Associated Adaptation

I find no significant evidence for classical models of polygenic adaptation for the genes in the micronutrient-associated gene sets. Rather, I suggest that micronutrient adaptation is likely often oligogenic in nature. In particular, this is likely for adaptation in response to selenium, calcium and zinc dietary levels given that these gene sets repeatedly show high numbers of genes exhibiting signatures of selection, but none at the level of documented cases of human polygenic adaptation (Pritchard et al. 2010; Daub et al. 2013; Berg and Coop 2014; Field et al. 2016; Berg, Harpak, et al. 2019; Berg, Zhang, et al. 2019) or across the majority of genes with a gene set (**Section 3.4.3**).

The limited signatures of polygenic adaptation are due to one of two reasons. The first is that there is indeed limited selection over each micronutrient gene set due to deleterious pleiotropy, which can limit the adaptative potential of some genes. Since many of the MA-

genes within each gene set have a multitude of roles surrounding not only micronutrient regulation (Monteiro et al. 2015), it is likely that such pleiotropy constrains polygenic adaptation within these gene sets.

In contrast, the limited signatures of polygenic adaptation observed may simply reflect the limitations of the methods used in this study. In **Chapter 2**, F_{ST} and $Relate$ are shown to have the highest power to identify local adaptation mediated by standing variation at both the monogenic and polygenic level (compared to haplotype-based methods to identify positive selection) but this power is still limited in some populations and for more recent selection. Indeed, using the best inferred method is not synonymous with identifying all signatures of positive selection, and I can only suggest that more true signatures of positive selection have been identified than if using such tested haplotype methods.

Still, it remains that the evidence for positive selection identified in the selenium, zinc and calcium gene sets, amongst other micronutrients, is dominated by strong signatures on only a few functionally-related genes. Hence, and in consideration of the above limitations, I propose that the term oligogenic adaptation is better suited when addressing adaptation to micronutrients in modern humans. This is as a generalisation of the observed signatures of positive selection across micronutrients and populations, and polygenicity of adaptation amongst MA-gene sets and populations is likely more intricate than can be fully appreciated in a study of this design.

3.5.3. Candidate Populations under Oligogenic Adaptation

I first identify populations as the most likely to have undergone micronutrient-associated adaptation under the assumptions of oligogenic adaptation (populations with MA-gene sets showing a higher number of genes exhibiting signatures of positive selection are more likely to have undergone adaptation in response to a micronutrient-associated selective force). This approach also allows the implicit comparison of the likelihood of micronutrient-associated adaptation amongst populations, as well as ranking the likelihood of adaptation to each micronutrient within each population. This is as a powerful method to consider natural selection amongst different populations and functionally-related gene sets.

Here I outline the MA-gene sets which show multiple signatures of positive selection amongst different genes in populations which live on soils with toxic or deficient levels of the micronutrient of interest. Discussion of the molybdenum gene set (which shows an excess of significant SNPs across a number of populations, isolated to two genes) is omitted here, given that the levels of molybdenum in soils and molybdenum deficiency/toxicity in humans is so poorly categorised.

There is an enrichment of signatures of positive selection amongst selenium-associated genes in East Asian populations (Xia et al. 2005). These signatures agree with previous studies, which have suggested polygenic or oligogenic adaptation in selenium metabolism in East Asian populations (White et al. 2015). I also identify enrichment of selenium-associated signatures of positive selection in African populations, who too live on selenium deficient soil, particularly in Malawi (Hurst et al. 2013; Ibrahim et al. 2019; Ligowe et al. 2020). Hence, this is in support of selenium-deficient soils driving signatures of positive selection in multiple selenium-associated genes, and suggest parallel adaptation to selenium-associated selective pressures amongst these two metapopulations.

The phosphorus-enriched signatures of positive selection of the American Pima population, who live on what is now Arizona, also co-occur with known phosphorus deficiencies in the local soil (or more accurately, co-occur with the low bioavailability of phosphorus from calcareous soil (Silvertooth et al. 2001; von Wandruszka 2006)). Salinity in African soils has been known to be highly variable and reaching both the deficient and toxic level (Nell and van Huyssteen 2018; Shahid et al. 2018; Hassani et al. 2021), potentially driving the observed adaptive signatures in various African populations. Still, it is difficult to confidently infer how much of contemporary soil deficiencies of either phosphorus or chloride is due to recent agricultural practices (see **Section 3.5.6**; (Nell and van Huyssteen 2018; Shahid et al. 2018; Dhaliwal et al. 2019; Hassani et al. 2021)).

The Maya population of America also show an unusual excess of signatures of positive selection in iodine-associated genes, inferred from the number of significant SNPs within this gene set and the number of individual genes with (top-ranking) signatures of iodine-associated positive selection. However, there is insufficient soil data to evaluate the if signatures of positive selection are supported by unusual iodine composition of local soils here. Still, iodine deficiency is prevalent in Mexico, which encompasses the region of the Maya population. In the modern Mexican population, the prevalence of goitre, the swelling of the thyroid gland caused by iodine deficiency, is at 54.6% (Hetzel and Nutrition 1988). However, no studies have been carried out to establish if there is a lower prevalence of goitre in the native Maya population (as would be expected under the proposed iodine-associated adaptation).

I have presented here the best examples of signatures of positive selection in MA-gene sets in populations supported by either unusual soil composition or endemic deficiencies of the associated micronutrient. Since the level of micronutrients across global soils, particularly at the level of individual population regions, is not comprehensively known, I only explore the potential support of some genetic signatures from corresponding soil micronutrient levels. Moreover, micronutrient-associated deficiencies can also result from general malnutrition and socio-economic status, and I am cautious in presenting this data as representing the selective drivers of putative adaptation. A more comprehensive understanding of local soil environment and susceptibility of micronutrient-associated disease by ancestry may reveal further informative correlations of genetic signatures and soil or disease within this dataset.

3.5.4. Candidate Populations under Monogenic Adaptation

Further candidate populations under micronutrient-associated adaptation were identified as those that show especially strong signatures of positive selection in individual MA-genes. 21 populations show MA-genes with evidence of selection at the most stringent threshold (multiple-testing threshold; $pvalue < 4.65e^{-6}$). This suggests that selection on individual micronutrient-associated genes can be strong, and corresponding allele frequency change very quick. This is not only similar to the many supported cases of monogenic adaptation to diet in humans (Tishkoff et al. 2007; Mathieson et al. 2015; Schlebusch et al. 2015; Minster et al. 2016; Mathieson and Mathieson 2018) but also in line with the assessment of limited polygenic adaptation.

I propose monogenic adaptation (or at least adaptation primarily mediated by one gene) in response to iron-associated pressures in two populations: the European Basque population (mediated by *HIF1A*) and the East Asian Yakut population (mediated by

FTMT). This agrees with previous studies suggesting iron-associated adaptation in various Eurasian populations but it remains unclear whether this is driven by the suggested soil levels, changes to the diet driven by the Neolithic transition or cold ambient temperatures (Distante et al. 2004; Ye et al. 2015; Heath et al. 2016b, 2016a).

Similarly, I propose magnesium-associated adaptation in two Central-South Asian populations, which show strong signatures of positive selection in two different magnesium-associated genes: *MECOM* in the Brahui and *FXYD2* in the Ugyur. Indeed, a mutation of the *FXYD2* gene has been linked to hypomagnesemia (Sha et al. 2008), potentially a consequence of adaptation to the well-categorised magnesium dominant soil of Central Asia (Vyshpolsky et al. 2008; Karimov et al. 2009). *MECOM* has not been explicitly linked to magnesium response, but has been associated with osteoporotic fractures (Hwang et al. 2013). Since magnesium is associated with bone density and prevents the onset of osteoporosis (Castiglioni et al. 2013), it is possible that this variant also confers hypomagnesemia.

Micronutrient-associated adaptation mediated by these genes may occur in populations other than those with the strongest signatures, but these are not presented here. Similarly, and analogous to that addressed in **Section 3.5.3**, many other genes bypassing the most stringent threshold have been identified, but only those with the strongest supporting evidence from soil data, functional role or surrounding literature are mentioned here. Still, the MA-genes bypassing this stringent threshold encompass nine micronutrient categories, and I suggest that strong local selection in response to micronutrient-associated selective pressures has indeed played a role in shaping human genetic variation.

3.5.5. Candidate Genes Mediating Widespread Adaptation

Whilst the signatures of positive selection do suggest that local, rather than more global, adaptation is more common in micronutrient-associated adaptation, some MA-genes show strong signatures (often bypassing the most stringent threshold) shared across multiple populations across the globe, and therefore exhibit evidence of widespread selection. This includes the zinc-associated genes *SLC39A11*, *SLC39A4* and *SLC30A9*; the phosphorus-associated gene *PDE7B*; the calcium-associated *ATP2B2*; and the *SLC12A1* gene associated with potassium, sodium and chloride metabolism. Widespread adaptation has previously been identified in zinc-transporter genes, including *SLC39A4* and *SLC30A9* (Zhang et al. 2015a; Engelken et al. 2016; Roca-Umbert et al. 2022), but adaptive signatures of the remaining genes in modern humans has not currently been recorded. Because of their potential importance in human dietary adaptation, I discuss each of these genes below.

PDE7B is a phosphodiesterase with variants associated with phosphorus serum levels (Kestenbaum et al. 2010) but primarily identified as playing a key role in cancer development (Cao et al. 2019; Sun et al. 2020) (and, interestingly, the silkiness of chicken feathers (Li et al. 2019)). Whilst the contemporary levels of phosphorus are heavily affected by agricultural practices (Dhaliwal et al. 2019; Alewell et al. 2020), there is a broad pattern of increased soil phosphorus in non-African environments, perhaps pertaining to this widespread signature (He et al. 2021).

The solute carrier gene *SLC12A1* is less easily associated with one particular micronutrient, since this gene mediates metabolism and transport of sodium, potassium and chloride. Still, in this study, its dominant micronutrient association is assigned as

sodium, since it accounts for much of the salt reabsorption in the kidneys (Markadieu and Delpire 2014). Mutations in *SLC12A1* result in Bartter's syndrome, which is an autosomal recessive disorder produced by the removal of too much salt from the body (Gagnon and Delpire 2013). Given that excess salt results in a significant increase of the risk of high blood pressure and associated co-morbidities (Hunter et al. 2022), it is possible that such mutations acting to remove salt may have been adaptive in environments characterised by hyper-saline soils, which are common at least in contemporary times (Nell and van Huyssteen 2018; Shahid et al. 2018; Hassani et al. 2021).

The plasma membrane calcium ATPase *ATP2B2* plays a key role in human health, associated with various cardiovascular diseases and deafness, amongst other conditions (Stafford et al. 2017). It also plays a critical role in intracellular calcium homeostasis and has also been associated with the calcium absorption pathways of laying hens (Gloux et al. 2019). However, given its association with multiple human diseases, it is unclear if the adaptive signatures observed here can be confidently associated with dietary calcium.

Since the signatures of positive selection of these genes are observed over most non-African populations, it is possible that the observed signatures of positive selection are a result of adaptation in an ancestral Out of Africa population. Indeed, soil in the Middle-East has been shown to be both zinc-deficient and hyper-saline (Ryan et al. 2013; Shahid et al. 2018), potentially driving the suggested widespread selection in the zinc-associated genes *SLC39A11*, *SLC39A4* and *SLC30A9* and the sodium-associated gene *SLC12A1*. Still, is also possible that other factors (such as different novel pathogens or temperatures increasing the selective pressure for pathogen-starvation or thermoregulation (Engelken et al. 2014; Pietrangelo 2015; Heath et al. 2016a)) drive these shared signatures of positive selection amongst non-Africans.

3.5.6. Summary

In summary, I show that signatures of positive selection associated with essential micronutrients exist in many geographic areas and multiple micronutrient categories. I also suggest that micronutrient-associated adaptation is primarily mediated by the genetic changes in a small number of micronutrient-associated genes. Known micronutrient soil levels support proposed adaptation of micronutrient-associated genes, and micronutrients may have played an important selective role in modern humans, potentially shaping the genomic variation of our species.

Still, the power of this comprehensive approach also coincides with some key limitations. This study is a broad overview into the nature of micronutrient-associated adaptation, but is not able to explore adaptation in specific micronutrient categories at depth. Here, I only outline the strongest signatures of positive selection in populations with additional support for micronutrient-associated adaptation (*i.e.,* from known soil composition, known endemic deficiencies or functional information of individual genes showcasing the signatures of positive selection), which biases the findings to this available data. Whilst impossible for a study of this design, individual and in-depth exploration of the signatures of positive selection identified for each micronutrient will more fully elucidate the role of individual micronutrients in modern human adaptation.

Moreover, contemporary soil and public health data cannot confidently be said to represent the ancestral micronutrient-associated selective pressures experienced by a population, since they are heavily impacted by modern agricultural practices and modern health inequality (Diamond 2002; Bhutta and Salam 2012; Bailey et al. 2015). The data

currently available can only evaluate some proposed links between soil as a selective pressure and proposed adaptation, and agreement between genetic signatures and micronutrient soil levels cannot be taken as conclusive evidence for adaptation.

It is clear that the signatures of positive selection identified within different micronutrient gene sets are highly variable, and likely represent a highly dynamic history of selection across these vital dietary components. This includes variable degrees of suggested polygenicity, variable geographic distribution of signatures of positive selection and variable genes suggested to mediate the same micronutrient-associated selective pressure between different localities. I therefore propose that adaptation in response to micronutrient-associated pressures present but limited in human evolutionary history, and highly complex across populations and micronutrient categories.

3.6. Conclusion

I infer the likely role of dietary micronutrients as a selective force across human populations and suggest that the adaptive responses to these selective forces have contributed to human genetic variation and population differentiation. I provide evidence that adaptation in response to micronutrients in the diet is most likely at the monogenic and oligogenic level. I show that in some cases the evidence of genetic adaptations is supported by local soil geology and suggest the micronutrients, including individual genes, with the strongest evidence of selection, and of which warrant further study.

Chapter 4: Evolutionary History of Micronutrient-Associated Genes in Modern Humans

4.1. Overview

Modern human populations have encountered a wide range of selective pressures over their history, often a result of environmental change, large-scale migrations into novel environments or cultural changes, some of which are linked to the diet (Perry et al. 2007; Tishkoff et al. 2007; Naugler 2008; Huerta-Sánchez et al. 2014; Fumagalli et al. 2015; Schlebusch et al. 2015; White et al. 2015; Engelken et al. 2016; Minster et al. 2016; Key et al. 2018; Rees et al. 2020). In previous work (**Chapter 3**), I infer that micronutrient levels in the diet have acted as a differential selective pressure across human populations and suggest that micronutrient-associated adaptation has contributed to shaping modern human genetic variation.

Here, I highlight five micronutrients (zinc, calcium, selenium, iron and iodine) which show particularly strong evidence for having evolved under natural selection across human populations, and across associated genes. I use a combination of signatures of positive selection that were previously identified (**Chapter 3**), gene network analysis, haplotype analysis and tree-based methods to explore the nature of putative selection events. This includes investigating the co-occurrence of signatures of positive selection on candidate genes across populations and inferring the most likely geographical origin and time of onset of proposed positive selection. Inferring the latter allows the suggestion of putative main selective drivers of micronutrient-associated adaptation, and how they may vary over populations and micronutrients.

I suggest that the same small groups of genes often mediate micronutrient-associated adaptation across populations, with additional genes further contributing to an adaptive response in some individual populations. However, I also identify outliers to this general trend, where different populations appear to mediate such adaptation via different groups of micronutrient-associated genes. From the geographical distribution of signatures of natural selection and the estimated timeframe or origin of positive selection, I propose that soil levels have largely driven adaptation in response to micronutrient levels, but identify potential examples of micronutrient-associated adaptation more closely surrounding the Neolithic transition, providing some evidence for the role of more recent dietary change in driving micronutrient-associated adaptation.

4.2. Background

Modern humans have been exposed to a plethora of selective pressures throughout their evolutionary history, particularly those pertaining to pathogens, diet and abiotic environmental factors like UV exposure or temperature (Perry et al. 2007; Tishkoff et al. 2007; Vernot and Akey 2014; Schlebusch et al. 2015; White et al. 2015; Minster et al. 2016; McManus et al. 2017; Key et al. 2018; Rees et al. 2020). Moreover, these selective pressures are often differentially exerted across populations, a by-product of our species unique colonisation of global and highly varied, sometimes extreme, environments (Ilardo and Nielsen 2018).

Such selective pressures can result in population-specific adaptations (Savolainen et al. 2013), which leave complex signatures of natural selection on modern genomes (Sabeti et al. 2006; Pritchard et al. 2010; Rees et al. 2020). Many studies have focused on identifying these signatures in populations across the globe (Sabeti et al. 2002; Ilardo and Nielsen 2018; Rees et al. 2020), resulting in a catalogue of suggested local adaptation targets in modern humans and, for some putative targets, their respective drivers of selection. Still, in many cases, the exact dynamics of selection, such as the onset or selective pressure, remain a question.

There has been some success in linking environmental or cultural pressures to identified signatures of natural selection across populations. For example, signatures of selection identified in genes associated with hypoxia resistance in populations living at high altitude or frequent diving (Ilardo et al. 2018), arsenic-resistance in an Argentinian population living on arsenic-toxic soil (Schlebusch et al. 2015) or temperature-sensation across populations living at northern latitudes (Key et al. 2018). However, in many cases, the selective pressures driving adaptation are still under debate. This includes the nature of the selective pressure(s) driving signatures of positive selection on genes associated with short stature in some rainforest populations, proposed to be driven by thermoregulatory pressures, locomotory advantages or a consequence of adaptation to iodine-deficiency (Herráez et al. 2009; Perry and Dominy 2009; Venkataraman et al. 2018); strong signatures identified in the *EDAR* gene in East Asian populations associated with hair thickness, tooth and ear shape, sweat gland density and chin protrusion, but with no clear selective driver (Sabeti et al. 2007; Fujimoto et al, 2008; Adhikari et al, 2015; Reyes-Reali *et al*, 2018, Speidel et al. 2019, Kataoka et al, 2021); and even the strong signatures upstream of the *LCT* conferring lactase persistence in European and African populations, long associated with the Neolithic transition but with frequency increases appearing to be considerably younger (Gerbault et al. 2011; Sverrisdóttir et al. 2014; Mathieson et al. 2015; Burger et al. 2020; Evershed et al. 2022).

4.2.1. Micronutrients as a Selective Pressure

Adaptation in response to micronutrient levels and metabolism has been identified in human populations in previous work (Engelken et al., 2014, 2016; Herráez et al., 2009; Kovacs et al., 2013; White et al., 2015; Ye et al., 2015; Zhang et al., 2015a), and in the work described in **Chapter 3**. This adaptation has been suggested to be driven by various factors, most commonly the content of micronutrients in local soil (thereby the levels being absorbed into the diet through consumed plant and animal matter) or cultural evolution of the diet.

Indeed, much of the current collection of work has proposed that the putative micronutrient-associated adaptation is driven by the micronutrient levels in local soil, particularly relating to selenium, zinc and iodine-metabolism in previous literature (Cifor 2006; Herráez et al. 2009; Hurst et al. 2013; White et al. 2015; Zhang et al. 2015a) and selenium, magnesium, phosphorus and chloride-metabolism in the work described in **Chapter 3**. Endemic pathologies, particularly when partnered with accompanying soil data, can also be assumed to be a by-product of insufficient soil concentrations or decreased bioavailability of trace minerals, and may also reflect a soil-related selective pressure (Cifor 2006; von Wandruszka 2006; Hurst et al. 2013), as is suggested in **Chapter 3** in regards to putative adaptation in response to iodine.

Still, the extent of the role that local soils have played in driving micronutrient-associated signatures of positive selection remains a question for many individual cases of putative micronutrient-associated adaptation. For example, previous studies have noted the widespread signatures of adaptation of zinc-transporter genes in European populations (Engelken et al. 2014; Zhang et al. 2015a; Roca-Umbert et al. 2022), which is also suggested in **Chapter 3.** It is currently unclear if the shared signatures of positive selection across European populations in zinc-transporter genes reflect convergent adaptation to many different soils inhabited by non-African populations, or an adaptation to a soil environment inhabited by a common non-African ancestral population. In the latter case, this soil environment may be that within the Middle-East, since these were the environments colonised by migrating human populations Out-of-Africa (Ryan et al. 2013).

Signatures of positive selection identified in micronutrient-associated genes may instead be driven by selective pressures other than micronutrient level (or more explicitly, bioavailability) in local soils. Cultural changes and differences in diets amongst populations may affect the levels of certain micronutrients consumed, and, in theory, impose selective pressures in maintaining optimum intake or metabolism of micronutrients. The Neolithic transition, beginning approximately 10kya, resulted in major changes in human societies, including those relating to food growth and acquisition (Dobrovolskaya 2005; Perry et al. 2007; Naugler 2008). In particular, the Neolithic transition and switch to agriculture has been associated with reduced iron and calcium in the human diet, amongst other key micronutrients (Diamond 2002; Naugler 2008; Gerbault et al. 2011). Agricultural practices also deplete soils of many micronutrients, including zinc, copper and iron (Diamond 2002; Dhaliwal et al. 2019), which may act as an additional driver of adaptation in micronutrient metabolism in more recent human history.

Other suggested selective drivers of micronutrient-associated adaptation include pathogen stress and temperature regulation (see **Section 1.7.3**). Hence, whilst many selective drivers have been suggested to explain proposed micronutrient-associated adaptation, there remains no clear consensus on many individual cases.

4.2.2. Timepoint and Polygenicity of Micronutrient-Associated Adaptation

Alongside pinpointing the exact selective driver behind proposed micronutrient-associated adaptation, the timepoint and polygenicity of micronutrient-associated adaptation should also be considered. In particular, the timepoint of positive selection acting on micronutrient-associated genes is particularly interesting for two main reasons: 1) no study has explicitly investigated the timepoint of proposed micronutrient-adaptation and 2) the timepoint of the onset of selection and the selective driver are intrinsically linked. If the onset of selection can be accurately inferred, this facilitates the identification of plausible selective drivers, such as micronutrient-deficient or toxic soils inhabited by a common non-African ancestral population (or other environmental stress, such as pathogen load, experienced by a common non-African ancestral population), more recently colonised soils (or more recently encountered environmental stress), or even more recent Neolithic changes to diet or agriculture.

The number and identity of genes that may contribute to adaptation in response to micronutrient levels is also an interesting and important question. Signatures of positive

selection do not stretch over the entire micronutrient gene sets, and I propose that the adaptation across many sets of micronutrient-associated genes is likely to be oligogenic rather than polygenic in nature (**Chapter 3**). Still, given the nature of the study in **Chapter 3**, this remains only a broad overview and remains to be fully investigated.

4.2.3. Focal Micronutrient-Associated Gene Sets

In this study, I focus on five individual micronutrient-associated gene sets, which allows a more in-depth analysis, and consequent understanding, of the signatures of positive selection identified on these micronutrient-associated genes. By choosing to explore adaptation in response to only five micronutrients, I am also more clearly able to compare the signatures of positive selection amongst different micronutrient-associated genes and contextualise the inferences.

I focus on the gene sets associated with zinc, calcium, selenium, iron and iodine, chosen for two main reasons (as informed by the work undertaken in **Chapter 3**). The first is that these gene sets show some of the strongest evidence of positive selection (according to the signatures of positive selection across multiple genes; signatures of positive selection in the same gene(s) shared across many populations; or signatures of positive selection in individual gene(s) which bypass the most stringent threshold). The second is that the signatures of positive selection in these micronutrient-associated gene sets vary in their geographic range, with some signatures of positive selection isolated in individual populations, where others are shared across continental groups. Hence, these gene sets demonstrate geographical breadth of signatures of positive selection and proposed selection. A final, additional reason is that, in comparison to other micronutrients, there is also relatively more data about the global soil concentrations of these micronutrients than others (*e.g.,* zinc, selenium, iron and iodine), which can provide supporting evidence to the claim of natural selection (Xia et al. 2005; Cifor 2006; Herráez et al. 2009; Hurst et al. 2013; Ryan et al. 2013).

These highlighted micronutrients are also particularly relevant to human health, with deficiencies of zinc, iodine and iron being the most common across the globe (25% of the world's population expected to be affected by either iron or iodine deficiency and 17% at risk from zinc deficiency (Bhutta and Salam 2012; Bailey et al. 2015; Khan et al. 2022)). Calcium and selenium deficiencies are also common, with dietary levels of calcium being estimated as deficient in approximately 50% of the world's population (Shlisky et al. 2022) and selenium deficiency affecting up to one billion people worldwide (Jones et al. 2017). In some populations, deficiencies of any of these five micronutrients are so common that they result in endemic pathologies, such as is the case with iron-associated anaemia and iodine-associated goitre (recorded across multiple global populations (Kelly and Snedden 1960; Dormitzer et al. 1989; Manning et al. 2012; Stevens et al. 2022)), and the cardiomyopathies and bone disorders recorded in selenium-deficient areas of China (Xia et al. 2005).

4.2.4. Study Overview

Using the results in **Chapter 3,** I first explore and compare the geographic distribution of the strongest signatures of positive selection in each micronutrient set and identify which genes have the strongest evidence of selection amongst global populations. I then explore and compare the proposed oligogenic adaptation in genes associated with zinc, calcium and selenium, and how the groups of genes that mediate micronutrient-associated

adaptation may differ across the globe. I also ask if the signatures of positive selection shared over many global populations, primarily in zinc, calcium and selenium-associated genes, are most likely driven by the same ancestral selective pressure. With this, I suggest which genes may have undergone adaptation swiftly following or surrounding the Out of Africa migration (Soares et al. 2012; Haber et al. 2019; Tucci and Akey 2019). Finally, I infer the most likely onset of selection for calcium and iron-associated genes and suggest whether changes in the diet related to the Neolithic transition (Dobrovolskaya 2005; Naugler 2008) or migrations into environments with varying soil levels have most likely driven this suggested selection.

The signatures of positive selection in zinc, selenium and calcium-associated genes form networks of only a few genes which are often shared by multiple individual populations, globally or within the same metapopulation. In contrast, the signatures of positive selection on iron and iodine-associated genes appear more unique to individual populations, and suggest that associated adaptation is more locally concentrated across populations. Finally, I suggest that the geographic and temporal origins of adaptation in response to micronutrient-levels are highly varied. Ultimately, I propose that both migrations into new environments, and corresponding novel soil composition, and recent agricultural and dietary change have played a role in shaping the adaptive response of micronutrient-associated genes across modern human populations.

4.3. Methods

4.3.1. Datasets

4.3.1.1. The Micronutrient-Associated Genes Dataset

I use gene-sets associated with the uptake, regulation and metabolism of the trace minerals selenium (n=61), zinc (n=46), iron (n=44), and iodine (n=18) and the macromineral calcium (n=23). The literature and databases used to curate these gene sets are described in **Section 3.3.1.** Zinc, calcium, selenium, iron and iodine-associated genes are hereafter referred to as ZCSII-associated genes. The abbreviation MA-genes (micronutrient-associated genes) and pMA-genes (proxy micronutrient genes, which act as the neutral background see **Section 3.3.1.2**) are also used in this chapter.

Following the application of a positive mask that removes segments of the genome of low reliability (see section **3.3.3** (Bergström et al. 2020)), 182 genes remain (176 of which are autosomal; see **Table S4.1**). Five genes are associated in the literature with two of these micronutrients: *SLC11A1* is associated with both iron and zinc; *DIO1, DIO2, DIO3* and *SECISBP2* are associated with both selenium and iodine.

I verify that the SNPs in these gene sets do not have a significantly different allele frequency distribution compared to the genomic background inferred from chr1 of the Yoruban individuals (see **Table S3.3** and **Section 3.3.1.1**). In terms of SNP density, five genes have high SNP density when compared to the generated pMA-gene regions (see **Section 3.3.1.2**, **Table S3.4**): *SELENOO* (selenium-associated), *EPAS1* (iron-associated, introgressed from Denisovans in East Asians (Huerta-Sánchez et al. 2014)), *MT1A* and *MT1F* (zinc-associated) and *SLC8A1* (calcium-associated). Finally, according to *Ensembl* (Yates et al. 2020), eleven of the ZCSII-genes overlap or are less than 10kbp apart: the zinc-associated genes *MT1F, MT1G* and *MT1H,* the zinc-associated pair *CA1* and *CA3*, and the selenium-associated pairs of genes of *LHFPL2* and *ARSB, DMGDH* and *BHMT2*, *GPx5*

and *GPx6* (see **Table S3.3.2**). Any signatures of positive selection in these overlapping gene regions are therefore treated as possible signatures for either gene region.

4.3.1.2. The Population Dataset

I use a dataset of 913 individuals from the HGDP dataset (as published by (Bergström et al. 2020)), including populations in Africa, the Middle-East, Europe, East Asia, Central-South Asia, Oceania and the Americas, as described in **Chapter 3**. These individuals are grouped into 40 populations (see **Table S3.6, Fig. 3.1**), either from the populations specified from (Bergström et al. 2020) or following population analysis and geographic proximity (see **Section 3.3.2, Fig. S3.3-9, S3.10-13**).

4.3.2. Methods to Identify Positive Selection

The methods to identify signatures of positive selection are identical to those described in **Chapter 3:** *Relate* (Speidel et al. 2019) and F_{ST} (as calculated for all population combinations with Yoruba, as well as for all population pairs within Africa (Weir and Cockerham 1984)). All information on these methods, pre-processing and filtering are given in **Section 3.3.3.**

As a brief summary, SNPs are extracted from the candidate genes (and their 10kb regions up- and downstream) which fall in the 0.1% tail of either the F_{ST} and *Relate* empirical genome-wide background and treat those SNPs as having evidence for selection. Here, I also identify SNPs which fall in the 0.01% tail of either the F_{ST} and *Relate* empirical distribution, and assign those SNPs as having strong evidence for selection of which to focus the analysis. Analogous to **Chapter 3,** I also extract SNPs which exhibit signatures of positive selection at the multiple-testing threshold of 4.65×10^{-6} used in **Section 3.4.5,** which is the most stringent threshold and identifies the SNPs with strongest evidence of selection in this study.

4.3.3. Gene Networks

Gene networks are built to identify which genes frequently share signatures of positive selection in the same populations. To do so, I first identify pairs of genes that share signatures in the 0.1% tail of the *Relate* empirical distribution for two or more populations. Here, I only consider signatures of positive selection according to *Relate* to avoid simply capturing groups of genes that are differentiated from the Yoruba population. Gene networks are then built using the *GGally* package in *R* (Schloerke et al. 2021), where genes are connected if they share signature of positive selection in two or more populations and the strength of the connection is proportional to the number of populations in which their signatures co-occur.

4.3.4. Haplotype Networks

Haplotype networks are built surrounding focal, candidate SNPs using *POPART* (Leigh and Bryant 2015). For genes that I identify as having strong evidence of positive selection, I assign focal SNPs as those with the evidence of positive selection shared over the highest number of populations or those representing regions of the candidate gene with clusters of signatures of positive selection (according to both F_{ST} and *Relate* evidence of selection; see **Table S4.2**). I choose to manually assign focal SNPs using this criterion, rather than a systematic approach, since any one criterion does not best represent the SNPs with the strongest evidence of selection across all genes of interest.

I then extract regions of 10kb and 20kb around these focal SNPs (masking genomic regions with low reliability (as inferred by (Bergström et al. 2020)), filtering for sites with MAF < 0.05, removing indels and only retaining biallelic sites) and phase these regions with $SHAPEIT2$ (Delaneau et al. 2013). The phased files are then reformatted to the required input file format of $POPART$ (Leigh and Bryant 2015), of which are used to build a median joining tree network.

4.3.5. Inferring Time of Selection

I infer the timing of selection on iron and calcium-associated genes to address the hypothesis that recent changes to the diet (*i.e.,* those surrounding the Neolithic transition) drove putative iron and calcium-associated adaptation. I first reconstruct the allele trajectories of focal SNPs of the iron and calcium-associated genes that are identified as having the strongest evidence of selection. Here, focal SNPs are identified as outlined in **Section 4.3.4** (see **Table S4.3**).

I estimate the onset of selection using two programmes: $Relate$ (Speidel et al. 2019) and $CLUES$ (Stern et al. 2019). $CLUES$ estimates the timing and strength of selection using a hidden Markov model, treating inferred local trees as the observed state and the allele frequency trajectory as the hidden state. Before using this programme, I reformat the inferred genealogies generated using $Relate$ into the $CLUES$ input format $newick$, which resembles the format of $ARGWEAVER$ (Rasmussen, Hubisz, et al. 2014). I then use $CLUES$ to infer allele frequency across time, and jointly estimate the strength and likelihood of selection beginning at 500, 1000, 1500 and 2000 generations ago (corresponding to 14kya, 28kya, 42kya and 56kya when using a generation time of 28 years (Speidel et al. 2019)) for each focal SNP. Here, log-likelihoods of over 4 are treated as moderate evidence of selection, in line with previous literature (Stern et al. 2019). I then use $Relate$ to trace the focal SNP's frequency across its lifetime and infer the timepoints surrounding striking frequency increases to evaluate the inferences from $CLUES$.

4.4. Results

4.4.1. Adaptive Signatures Across Micronutrients

Since there is limited evidence of polygenic adaptation (see **Chapter 3**), I first explore the geographical distribution of the strong signatures of positive selection on individual genes for each MA-gene set. In the following sections, the geographic distribution of the signatures of positive selection over each ZSCII-associated gene set are briefly recapped and compared (see **Fig. 4.1, Fig. 4.2**; full lists of genes with signatures in the 0.1% tail of the empirical distributions of F_{ST} and $Relate$ are given in **Tables S4.3-13**). I then verify if stronger signatures of positive selection maintain this geographic distribution. To do so, I identify SNPs that are in the 0.01% tail of the empirical distributions of F_{ST} and $Relate$ (*i.e.,* using a threshold that is a magnitude more stringent), as well as those with signatures at the most stringent threshold as used in **Chapter 3** (4.65×10^{-6} ; see **Section 3.4.5**).

The strongest signatures of positive selection are discussed, including what they suggest regarding the degree of polygenicity and geographic range of proposed adaptation, for each micronutrient gene set below. I also consider if the strongest signatures of positive selection isolated to a small number of populations are likely to be truly representing ultra-local selection events, or if there is only power to identify the signatures of positive

selection in certain populations. To evaluate this, I ask if the genes with the strongest signatures of positive selection have signatures in the less stringent 1% tail of the empirical *Relate* distributions in other populations. Here, I only explore the signatures of positive selection inferred by *Relate* since the signatures inferred by F_{ST} simply identify genes that are highly differentiated from the Yoruba population, and therefore are less geographically informative.

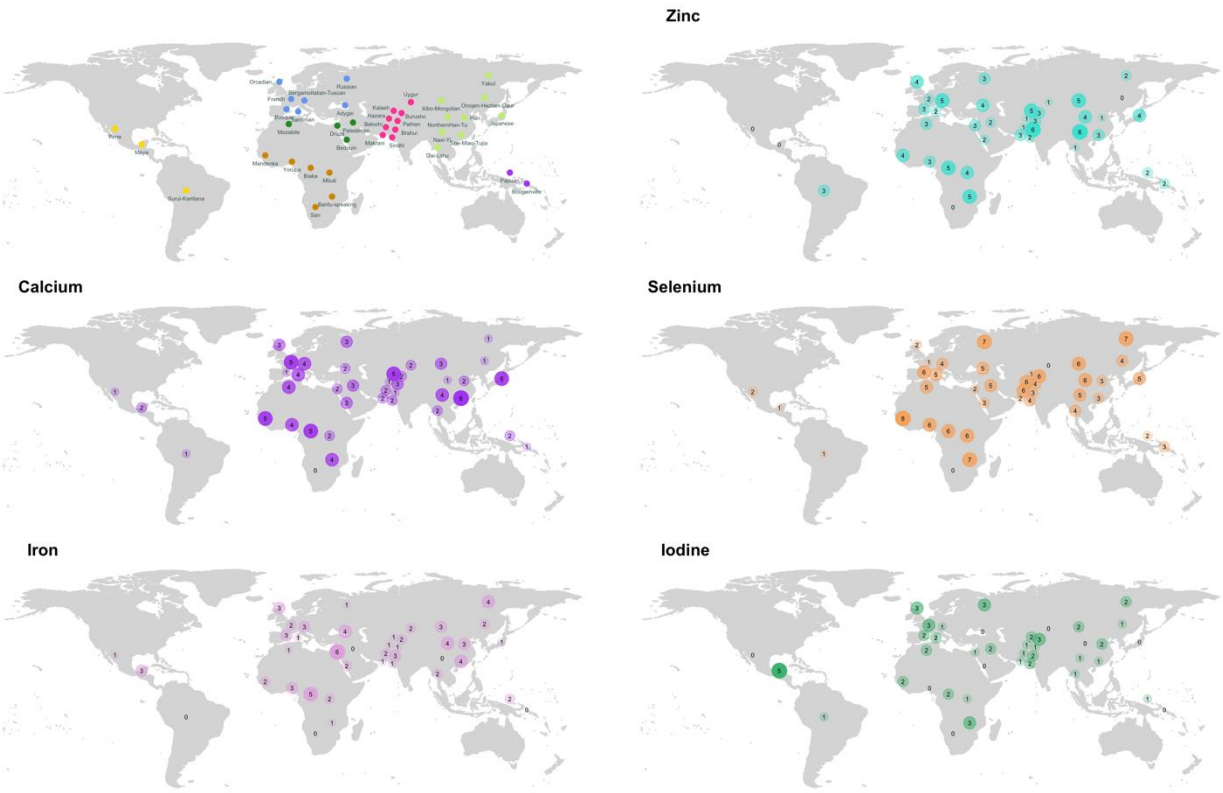


Fig. 4.1: Number of ZCSII-genes with Relate signatures of positive selection. *Signatures of positive selection (SNPs in the 0.1% tail of the Relate empirical distribution) for each population, given separately for genes associated with zinc, calcium, selenium, iron and iodine. Population names are given in the top left map.*

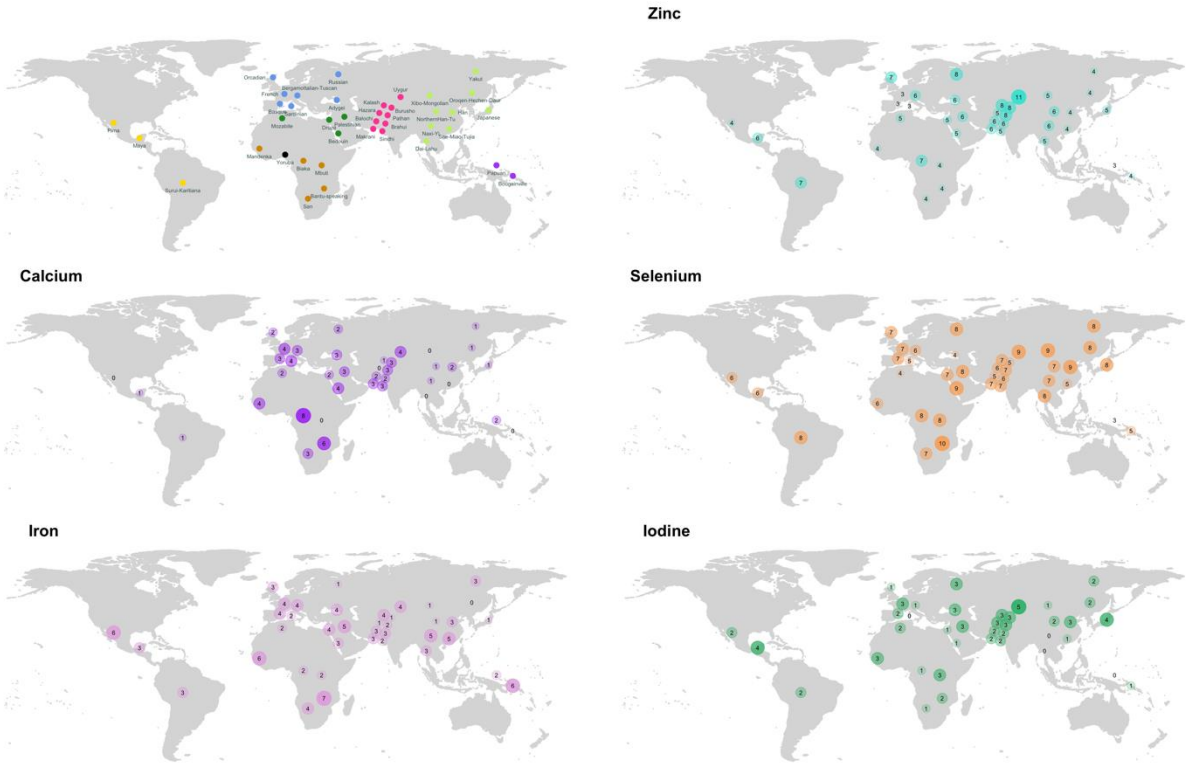


Fig. 4.2: Number of ZCSII-genes with F_{ST} signatures of positive selection. *Signatures of positive selection (SNPs in the 0.1% tail of the F_{ST} empirical distribution) for each population, given separately for genes associated with zinc, calcium, selenium, iron and iodine. Population names are given in the top left map.*

4.4.1.1. Zinc

In comparison to all other ZCSII-associated gene sets, the zinc-associated gene set (n=46) shows the highest number of genes with signatures of positive selection shared amongst many populations (**Tables S4.4-5**)). Of these zinc-associated genes, 16 show strong signatures of selection (**Table 4.1**). Many of these genes are zinc-transporters (*e.g., SLC39A4, SLC30A9, SLC39A11, SLC39A12, SLC30A7, SLC30A8, SLC30A1, SLC39A10, SLC39A14* and *SLC30A10,* see **Table 4.1**), with three exhibiting significant signatures at the most stringent threshold ($pvalue < 4.65 \times 10^{-6}$). *GPR39,* not a zinc-transporter gene, but associated with zinc-dependent signalling, also exhibits signatures of positive selection in all metapopulations bar Oceania, hence also presenting strong evidence for selection (**Table 4.1**).

For some of these candidate genes, the signatures of positive selection appear to be isolated to only one population, suggesting that either selection is ultra-local, or the thresholds are so stringent that nearly-significant signatures of positive selection in nearby populations are missed. When I consider those nearly-significant signatures (those in the 1% tail of the empirical distribution of *Relate*), all zinc-transporter genes

highlighted here (none of which are genomic neighbours) have nearly-significant or significant signatures of positive selection in more than 12 populations (**Table 4.2**). Given that these significant or nearly-significant signatures of positive selection are shared amongst many populations, and are observed in a functionally-related gene set previously shown to have an excess of significant SNPs according to both $Relate$ and F_{ST} (**Chapter 3**), I therefore suggest that this is indicative of widespread adaptation in response to a zinc-associated selective pressure.

I now consider the genes with strong signatures of positive selection, inferred either from $Relate$ or F_{ST}, across many populations (**Table 4.1**), and are therefore the strongest candidates for widespread adaptation: *SLC39A4, SLC30A9, SLC39A11* and *GPR39.* For three of these genes, *SLC39A4, SLC30A9* and *GPR39,* the widespread strong signatures of positive selection are inferred from F_{ST} (*SLC39A4* across almost all Eurasian populations; *SLC30A9* across almost all East Asian and some Central-South Asian populations; *GPR39* across many European and Central-South Asian populations (**Table 4.4.1**). In two instances, the evidence of positive selection bypasses the most stringent threshold ($pvalue < 4.65 \times 10^{-6}$) : *SLC39A4* in the Makrani ($F_{ST}\ pvalue = 3.95 \times 10^{-6}$) and *SLC30A9* in the Han ($F_{ST}\ pvalue = 3.55e \times 10^{-6}$).

There is no evidence of positive selection for *SLC39A4, SLC30A9* or *GPR39* within Yoruba as inferred by $Relate$ ($pvalues$ are not within the 0.1% or 1% tail of the empirical distribution), and therefore the differentiation from Yoruba, as captured by F_{ST} , is unlikely to represent selection within the Yoruba population. There are also no significant signatures of positive selection inferred for any of these three genes in any other African populations according to $Relate$, with the exception of *GPR39* identified within the 0.1% tail of the empirical distribution in the Mbuti population. Hence, I suggest that that the strong signatures of positive selection identified in these particular zinc-associated genes, which reach the most stringent threshold in two instances, are most likely a signal of Out-of-Africa positive selection in response to zinc.

The final zinc-associated candidate for widespread adaptation, *SLC39A11*, has strong signatures of positive selection amongst many populations, as inferred using $Relate$ rather than F_{ST}. Again, the strongest signature is captured in the Makrani population ($pvalue = 1 \times 10^{-6}$), but there are significant and nearly-significant signatures of positive selection amongst the majority of populations (see **Table 4.2**). I also then suggest that the signature of positive selection identified in the Makrani is not indicative to an ultra-local selection event, but is simply the strongest identified signature of a widespread selection event.

Hence, I suggest that zinc-associated adaptation, primarily mediated by the zinc-transporter genes given in **Table 4.1**, is most likely a result of widespread selection. This is most likely a selection event in the ancestors of non-Africans, and present *SLC39A4, SLC30A9, SLC39A11* and *GPR39* as the strongest candidates for mediating such proposed zinc-associated adaptation. Moreover, given that many zinc-associated genes show significant or nearly-significant signatures of positive selection within populations (most clearly observed in Russian, Uygur, Kalash and Burusho populations, see **Fig. 5.2, Table S4.2**), I also suggest that zinc-adaptation is oligogenic in nature, and that multiple genes may be involved in mediating zinc-associated selective pressures.

***Table 4.1**: **Zinc-associated genes with p-values** $< \mathbf{10^{-5}}$**.** P-values as calculated from the empirical distribution of either Relate or F_{ST}. P-values less than 4.65×10^{-6} (see **Section 3.4.5**) highlighted in bold.*

Gene	Population	*Relate* Significance	F_{ST} Significance
SLC39A4	San		9.96e-5
	Druze		1.02e-5
	Palestinian		2.85e-5
	Adygei		2.30e-5
	Basque		1.41e-5
	BergamoItalian-Tuscan		8.84e-6
	French		5.58e-6
	Orcadian		3.19e-5
	Russian		6.43e-6
	Sardinian		1.33e-5
	Balochi		2.89e-5
	Brahui		7.98e-6
	Burusho		8.76e--6
	Hazara		7.24e-6
	Kalash		1.72e-5
	Makrani		**3.95e-6**
	Pathan		5.47e-6
	Sindhi		1.03e-5
	Dai-Lahu		7.82e-5
	Han		4.34e-5
	Japanese		6.69e-5
	Oroqen-Hezhen-Daur		3.99e-5
	Naxi-Yi		9.03e-5
	NorthernHan-Tu		5.98e-5
	She-Miao-Tujia		4.75e-5
	Xibo-Mongolian		5.49e-5
	Yakut		3.57e-5
GPR39	Mbuti		9.97e-5
	Bedouin		5.59e-5
	Druze		3.23e05
	Palestinian		3.72e-5
	BergamoItalian-Tuscan		8.94e-5
	French		5.35e-5
	Russian		6.47e-5
	Sardinian		5.57e-5
	Brahui		2.63e-5
	Burusho		5.78e-5
	Hazara		9.24e-5
	Pathan		4.37e-5
	Sindhi		5.83e-5
	Japanese	7.5e-5	
SLC30A9	Bantu-speaking		2.83e-5
	Burusho		5.38e-5
	Hazara		2.23e-5
	Pathan		6.08e-5
	Dai-Lahu		2.12e-5
	Han		**3.55e-6**
	Oroqen-Hezhen-Daur		1.51e-5
	NorthernHan-Tu		7.01e-5

	She-Miao-Tujia		2.05e-5
	Xibo-Mongolian		2.66e-5
	Yakut		1.71e-5
	Maya		8.30-5
SLC39A11	Bantu-speaking	3.33e-5	
	Palestinian	1.48e-5	
	French	9.13e-5	
	Balochi	1.55e-5	
	Kalash		9.98e-5
	Makrani	**1.4e-6**	
	Sindhi		2.99e-5
	Naxi-Yi	8.43e-5	
	NorthernHan-Tu	3.90e-5	
	She-Miao-Tujia	2.41e-5	
SLC39A12	Mandenka		4.04e-5
	Makrani	2.4e-5	
SLC30A7	Bantu-speaking	8.8e-5	
	Pathan	9.46e-5	
SCAMP5	Yoruba	4.17e-5	
MTF1	Mandenka	1.04e-5	
CA1	**Mozabite**	**7.15e-6**	
SLC30A8	BergamoItalian-Tuscan	4.48e-5	
SLC30A1	Russian	1.32e-5	
SLC39A10	Kalash	1.21e-5	
CAR13	Bantu-speaking		8.23e05
SLC39A14	Palestinian		5.81e-5
SLC30A10	Orcadian		9.97e-5
MTF2	Hazara		9.77e-5

***Table 4.2**: **The number ("No.") and name of the populations ("Populations") with signatures of positive selection**. Signatures of positive selection as identified by the 1% or 0.1% tail of the empirical background distribution of Relate, for all zinc-associated genes with p-values $< 10^{-5}$ in at least one population (**Table 4.1**)*

Gene	1% tail		0.1% tail	
	No.	**Populations**	**No.**	**Populations**
SLC39A4	6	Mandenka, Mozabite, Palestinian, Bedouin, BergamoItalian-Tuscan, Maya	0	
SLC30A9	18	Palestinian, Druze, Adygei, BergamoItalian-Tuscan, French, Orcadian, Russian, Makrani, Sindhi, Brahui, Pathan, Burusho, Kalash, Uygur, Xibo-Mongolian, Yakut, Maya, Papuan	7	Orcadian, Makrani, Sindhi, Brahui, Pathan, Uygur, Papuan
GPR39	21	San, Mbuti, Biaka, Mozabite, Palestinian, Bedouin, Sardinian, Basque, Russian, Sindhi, Balochi, Pathan, Yakut, Japanese, Han, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Surui-Karitiana, Papuan, Bougainville	5	Mbuti, Mozabite, Japanese, She-Miao-Tujia, Dai-Lahu

SLC39A11	33	San, Bantu-speaking, Mbuti, Biaka, Yoruba, Mozabite, Palestinian, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, French, Orcadian, Russian, Makrani, Sindhi, Balochi, Brahui, Hazara, Burusho, Kalash, Xibo-Mongolian, Yakut, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Maya, Surui-Karitiana, Bougainville	27	Bantu-speaking, Biaka, Palestinian, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, French, Orcadian, Russian, Makrani, Sindhi, Balochi, Brahui, Hazara, Burusho, Kalash, Xibo-Mongolian, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Surui-Karitiana, Bougainville
SLC39A12	23	San, Bantu-speaking, Palestinian, Bedouin, Adygei, BergamoItalian-Tuscan, Basque, French, Orcadian, Russian, Makrani, Sindhi, Brahui, Pathan, Kalash, Yakut, Japanese, Han, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Maya, Bougainville	3	BergamoItalian-Tuscan, Makrani, Brahui
SLC30A7	15	Bantu-speaking, Mandenka, Bedouin, Hazara, Pathan, Burusho, Uygur, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, Han, NorthernHan-Tu, Naxi-Yi, Maya	3	Bantu-speaking, Bedouin, Pathan
SCAMP5	5	Yoruba, Orcadian, Pathan, Yakut, She-Miao-Tujia	1	Yoruba
MTF1	6	San, Mandenka, Palestinian, Oroqen-Hezhen-Daur, Japanese, Han	1	Mandenka
CA1	7	Bantu-speaking, Biaka, Mozabite, Palestinian, Druze, Basque, Papuan	1	Mozabite
SLC30A8	22	San, Bantu-speaking, Biaka, Yoruba, Mandenka, Mozabite, Palestinian, Adygei, BergamoItalian-Tuscan, French, Russian, Balochi, Pathan, Kalash, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, She-Miao-Tujia, Naxi-Yi, Pima, Surui-Karitiana	8	Bantu-speaking, Yoruba, Mozabite, Adygei, BergamoItalian-Tuscan, French, Kalash, Japanese
SLC30A1	25	San, Bantu-speaking, Biaka, Yoruba, Mandenka, Palestinian, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, French, Orcadian, Russian, Makrani, Sindhi, Balochi, Brahui, Hazara, Burusho, Kalash, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Maya, PapuanHighlands_PapuanSepi	8	Biaka, Adygei, Basque, Orcadian, Burusho, Kalash, NorthernHan-Tu, Naxi-Yi
SLC39A10	12	Biaka, Bedouin, BergamoItalian-Tuscan, Sardinian, French, Russian, Sindhi, Balochi, Brahui, Kalash, Naxi-Yi, Surui-Karitiana	2	Kalash, Surui-Karitiana
CAR13	4	Biaka, Palestinian, Druze, Papuan	1	Druze
SLC39A14	13	San, Yoruba, Mandenka, Mozabite, Druze, Bedouin, BergamoItalian-Tuscan, Makrani, Brahui, Kalash, Japanese, Naxi-Yi, Pima	1	Brahui
SLC30A10	24	San, Bantu-speaking, Biaka, Yoruba, Mandenka, Palestinian, Adygei, BergamoItalian-Tuscan, Basque, French,	9	Biaka, Adygei, Basque, Orcadian, Russian, Burusho, Kalash, NorthernHan-Tu, Naxi-Yi

		Orcadian, Russian, Makrani, Sindhi, Balochi, Brahui, Hazara, Burusho, Kalash, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Maya, Papuan		
MTF2	14	Yoruba, Mandenka, Bedouin, Sardinian, Russian, Sindhi, Balochi, Hazara, Kalash, Xibo-Mongolian, Oroqen-Hezhen-Daur, Han, NorthernHan-Tu, She-Miao-Tujia	3	Mandenka, Xibo-Mongolian, NorthernHan-Tu

4.4.1.2. Calcium

Similar to the zinc gene set (n=46), the calcium gene set (n=23) also contains genes with signatures of positive selection shared over populations spanning each major global area (**Table 4.3**). However, in comparison to zinc-associated genes, the signatures of positive selection over calcium-associated genes are 1) shared over fewer populations and 2) more frequently inferred by *Relate* rather than F_{ST}. Hence, in comparison to zinc, there is not the same preliminary evidence for selection on an ancestral non-African population, which is now explored.

ATP2B2 and *SLC8A1* show evidence for positive selection over the most populations. Of these two genes, *ATP2B2* shows the strongest evidence of selection, with signatures of positive selection identified at the most stringent threshold ($pvalue < 4.65 \times 10^{-6}$) in two populations (Mandenka; F_{ST} $pvalue = 7.75 \times 10^{-8}$; Sardinian; $Relate\ pvalue = 2.1 \times 10^{-7}$). Moreover, the strong signatures of positive selection in *ATP2B2* are observed in eleven populations spanning all metapopulations bar East Asia (**Table 4.3**) and significant and nearly-significant signatures of positive selection, as inferred by *Relate*, are observed in 33 populations, including five African populations. Therefore, *ATP2B2* is a strong candidate gene for near-global selection in modern humans, possibly responding to calcium-associated selective pressures.

There does not appear to be an excess of strong differentiation to Yoruba, as calculated by F_{ST}, at the gene set level in Eurasia (**Fig 4.2**). Further, there are more population-specific signatures of positive selection at the gene set level, as captured by either *Relate* or F_{ST} , than in zinc-related genes (**Fig. 4.1-2; Table S4.6-7**). In particular, many calcium-associated genes exhibit evidence of selection in the Biaka and Bantu-speaking populations of Africa, the She-Miao-Tujia and Japanese of East Asia, the Kalash of Central-South Asia and the French of Europe (**Fig. 4.1-2**). Hence, whilst some calcium-associated genes, *e.g., ATP2B2*, may have undergone widespread adaptation, it appears that oligogenic adaptation in response to calcium levels is less widespread, and may only be present in a few independent populations.

***Table 4.3**: **Calcium-associated genes with p-values** $< \mathbf{10^{-5}}$. P-values calculated from the empirical distribution of either Relate or F_{ST}. P-values less than* 4.65×10^{-6} *(see **Section 3.4.5**) highlighted in bold.*

Gene	Population	*Relate* Significance	F_{ST} Significance
ATP2B2	**Mandenka**		**7.75e-8**
	Bedouin		8.84e-5
	Mozabite	1.38e-5	9.73e-5
	French	7.51e-5	7.13e-5
	Sardinian	**2.1e-7**	
	Burusho	9.35e-5	
	Makrani	4.97e-5	
	Pathan	2.08e-5	
	Uygur	5.23e-5	
	Pima	9.06e-6	
	Papuan	1.97e-5	
SLC8A1	Biaka		3.05e-5
	Mandenka		3e-5
	BergamoItalian-Tuscan	4.48e-5	
	Makrani	2.4e-5	
	Dai-Lahu	7.39e-5	
	Maya	4.17e-5	
	Papuan	1.26e-5	
SLC8A3	Dai-Lahu	4.43e-5	
	NorthernHan-Tu	3.9e-5	
	Xibo-Mongolian	6.74e-5	
KCNJ10	Bantu-speaking		1.38e-5
	Mandenka		3.01e-5
ATP2B4	Biaka		8.39e-5
	Mozabite		8.32e-5
DGKD	Biaka	1.97e-5	
SLC12A3	Kalash	2.21e-5	

***Table 4.4**: **The number ("No.") and name of the populations ("Populations") with signatures of positive selection**. Signatures of positive selection identified by the 1% or 0.1% tail of the empirical background distribution of Relate, for all calcium-associated genes with p-values* $< 10^{-5}$ *in at least one population (**Table 4.3**)*

Gene	1% tail		0.1% tail	
	No.	Populations	No.	Populations
ATP2BP2	6	San, Mbuti, Biaka, Yoruba, Mandenka, Mozabite, Palestinian, Bedouin, BergamoItalian-Tuscan, Sardinian, Basque, French, Orcadian, Russian, Makrani, Brahui, Hazara, Pathan, Burusho, Kalash, Uygur, Xibo-Mongolian, Oroqen-Hezhen-Daur, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Pima, Surui-Karitiana, Papuan, Bougainville	8	Mozabite, Druze, Bedouin, BergamoItalian-Tuscan, Sardinian, French, Sindhi, She-Miao-Tujia

SLC8A1	32	San, Bantu-speaking, Mbuti, Biaka, Yoruba, Mandenka, Palestinian, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, French, Russian, Makrani, Balochi, Hazara, Pathan, Burusho, Kalash, Uygur, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Maya, Surui-Karitiana, Papuan	26	Bantu-speaking, Biaka, Yoruba, Mandenka, Palestinian, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, French, Russian, Makrani, Balochi, Hazara, Pathan, Kalash, Uygur, Yakut, Japanese, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Maya, Surui-Karitiana, Papuan
SLC8A3	22	Biaka, Yoruba, Mandenka, Palestinian, Druze, BergamoItalian-Tuscan, Sardinian, Orcadian, Sindhi, Balochi, Brahui, Pathan, Burusho, Kalash, Xibo-Mongolian, Yakut, Japanese, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Surui-Karitiana	10	Mandenka, Pathan, Burusho, Kalash, Xibo-Mongolian, Japanese, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu
KCNJ10	8	Mbuti, Mandenka, Palestinian, Druze, French, Brahui, Burusho, Naxi-Yi	0	
ATP2B4	14	Biaka, Yoruba, Mozabite, Druze, Bedouin, BergamoItalian-Tuscan, Sardinian, French, Sindhi, Hazara, Pathan, Xibo-Mongolian, She-Miao-Tujia, Pima	8	Mozabite, Druze, Bedouin, BergamoItalian-Tuscan, Sardinian, French, Sindhi, She-Miao-Tujia
DGKD	19	San, Bantu-speaking, Biaka, Yoruba, Mozabite, Palestinian, Bedouin, Orcadian, Russian, Makrani, Sindhi, Brahui, Burusho, Kalash, Oroqen-Hezhen-Daur, Yakut, Japanese, Naxi-Yi, Maya	5	Biaka, Orcadian, Russian, Kalash, Japanese
SLC12A3	10	Mandenka, BergamoItalian-Tuscan, Sardinian, French, Russian, Burusho, Kalash, Xibo-Mongolian, Yakut, She-Miao-Tujia	1	Kalash

4.4.1.3. Selenium

The selenium gene set (n=61) also contains genes with signatures of positive selection which are frequently shared across populations of every major global region (**Table S4.8-9**), but the individual evidence for positive selection in these genes is often weaker than that observed in zinc or calcium-associated genes. Many signatures of positive selection are identified according to the *pvalue* of 0.001 for either *Relate* or F_{ST}, but many fail to reach the *pvalue* threshold of 0.0001 (indicating strong signatures of positive selection; **Table 4.5**), and none reach the most stringent threshold of 4.65×10^{-6}.

African and East Asian populations often have the highest number of selenium-associated genes exhibiting evidence of positive selection (**Fig. 4.1-2**), consistent with a model of oligogenic selenium-associated adaptation in these regions. Still, East Asian populations do not appear to have a particular excess of strong signatures of positive selection when compared with European populations (**Table 4.5**), with the exception of those observed in *PRKG1*. Hence, individual genes may largely mediate selenium-associated adaptation (*e.g., PRKG1*), but many other additional genes, exhibiting weaker signatures of positive selection, may also be involved in the adaptative process (in agreement with the polygenic or oligogenic adaptation of selenium metabolism suggested in (White et al. 2015)).

The strongest signatures of positive selection in selenium-associated genes differ between African and East Asian populations: *PRKG1* shows signatures of positive selection in East Asian populations whereas *LRP8* and *LHFPL2* only carry strong signatures (at the 0.0001 *pvalue* threshold) in African populations (**Table 4.5**). This is consistent with different genes mediating adaptation to selenium across these different metapopulations. However, there are significant and nearly-significant signatures of positive selection, as inferred by *Relate*, in *PRKG1, LRP8* and *LHFPL2* in both East Asian and African populations (**Table 4.6**). This may indicate that the same groups of genes mediate adaptation in response to selenium levels, but the genes that primarily mediate this adaptation may differ between metapopulations.

***Table 4.5**: **Selenium-associated genes with p-values** $< \mathbf{10^{-5}}$**.** P-values calculated from the empirical distribution of either Relate or F_{ST}. P-values less than 4.65×10^{-6} (see **Section 3.4.5**) highlighted in bold.*

Gene	Population	*Relate* Significance	F_{ST} Significance
PRKG1	Bantu-speaking		2.11e-5
	Palestinian	6.83e-5	
	Hazara	5.56e-5	
	Han	6.54e-5	
	Naxi-Yi	2.17e-5	
	She-Miao-Tujia		2.05e-5
	Xibo-Mongolian		1e-5
SGCD	Biaka	8.77e-5	
	Mbuti	8.72e-5	
	Palestinian	9.50e-5	
	BergamoItalian-Tuscan		1.07e-5
	Makrani		1.18e-5
	Papuan		5.36e-5
AKAP6	Mozabite	4.16e-5	
	Adygei	4.62e-5	
	Yakut	7.62e-5	
	Surui-Karitiana	3.96e-5	
EEFSEC	Bantu-speaking		2.53e-5
	Bedouin		8.33e-5
	Mozabite		1.29e-5
	Basque	5.59e-5	
LHFPL2	Bantu-speaking		4.99e-6
	Biaka		8.33e-5
	Mbuti		7.44e-5
	San		3.08e-5
LRP8	Bantu-speaking	8.8e-5	
	Mandenka	1.04e-5	
	San	2.15e-5	
SELENOS	Russian		4.01e-5
	Brahui		8.04e-5
	Hazara		1.86e-5
KCNMA1	NorthernHan-Tu	3.9e-5	
	Yakut	1.46e-5	
SLCY	Mozabite	9.09e-5	
TXNDR3	Basque	4.79e-5	

SECISBP2	Sardinian	3.27e-5	
AKR7L	Burusho	9.83e-6	
GPx2	Makrani	9.61e-6	
TXNRD2	Mandenka		9.72e-5
TRU-TCA2-1	Mbuti		6.63e-5
ARSB	Orcadian		7.92e-5
SELENOM	Yoruba	5.87e-6	
SELENOP	Pathan	4.09e-5	
SELENOW	Japanese		6.91e-5
SEPHS2	Xibo-Mongolian		2.66e-5

***Table 4.6**: **The number ("No.") and name of the populations ("Populations") with signatures of positive selection**. Signatures of positive selection identified by the 1% or 0.1% tail of the empirical background distribution of Relate, for all selenium-associated genes with p-values $< 10^{-5}$ in at least one population (**Table 4.5**)*

Gene	1% tail		0.1% tail	
	No.	Populations	No.	Populations
PRKG1	30	Bantu-speaking, Biaka, Yoruba, Mandenka, Palestinian, Bedouin, Adygei, Sardinian, Basque, French, Orcadian, Russian, Sindhi, Balochi, Brahui, Hazara, Pathan, Burusho, Kalash, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Surui-Karitiana, Papuan	24	Bantu-speaking, Biaka, Yoruba, Mandenka, Palestinian, Bedouin, Adygei, Sardinian, Basque, Sindhi, Balochi, Hazara, Pathan, Burusho, Kalash, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu
SGCD	29	Bantu-speaking, Mbuti, Biaka, Mozabite, Palestinian, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, French, Sindhi, Hazara, Pathan, Burusho, Kalash, Uygur, Oroqen-Hezhen-Daur, Yakut, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Pima, Papuan, Bougainville	21	Bantu-speaking, Mbuti, Biaka, Mozabite, Palestinian, Druze, Bedouin, BergamoItalian-Tuscan, Sardinian, Basque, Sindhi, Hazara, Burusho, Oroqen-Hezhen-Daur, Yakut, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Bougainville
AKAP6	25	Bantu-speaking, Biaka, Yoruba, Mandenka, Mozabite, Palestinian, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, Orcadian, Sindhi, Pathan, Burusho, Uygur, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, NorthernHan-Tu, NorthernHan-Tu, Naxi-Yi, Maya, Surui-Karitiana, Papuan	20	Bantu-speaking, Biaka, Yoruba, Mandenka, Mozabite, Palestinian, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, Burusho, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, NorthernHan-Tu, Naxi-Yi, Surui-Karitiana, Papuan
EEFSEC	18	San, Bantu-speaking, Mbuti, Biaka, Yoruba, Mozabite, Palestinian, Bedouin, Adygei, Basque, French, Makrani, Sindhi, Balochi, Burusho, She-Miao-Tujia, Dai-Lahu, Surui-Karitiana	5	Bantu-speaking, Mbuti, Adygei, Basque, Balochi
LHFPL2	28	Mbuti, Biaka, Yoruba, Mandenka, Mozabite, Palestinian, Bedouin, Adygei, BergamoItalian-Tuscan, Basque, French,	4	Biaka, Mandenka, Japanese, Maya

		Orcadian, Russian, Makrani, Sindhi, Balochi, Brahui, Hazara, Pathan, Uygur, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Maya		
LRP8	15	Bantu-speaking, Yoruba, Mandenka, Mozabite, Russian, Makrani, Sindhi, Balochi, Brahui, Burusho, Xibo-Mongolian, Japanese, She-Miao-Tujia, Dai-Lahu, Papuan	5	Bantu-speaking, Yoruba, Mandenka, Balochi, Xibo-Mongolian
SELENOS	15	San, Biaka, Yoruba, Mandenka, Palestinian, Basque, Russian, Balochi, Burusho, Uygur, Xibo-Mongolian, Oroqen-Hezhen-Daur, NorthernHan-Tu, Dai-Lahu, Pima	5	Biaka, Yoruba, Palestinian, Balochi, NorthernHan-Tu
KCNMA1	23	San, Mbuti, Biaka, Yoruba, Mandenka, Adygei, BergamoItalian-Tuscan, Sardinian, French, Russian, Makrani, Sindhi, Burusho, Xibo-Mongolian, Yakut, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Pima, Bougainville	16	Mbuti, Biaka, Yoruba, Mandenka, Adygei, BergamoItalian-Tuscan, Russian, Sindhi, Xibo-Mongolian, Yakut, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu
SCLY	8	Bantu-speaking, Biaka, Mozabite, Palestinian, Bedouin, French, Russian, Brahui	4	Bantu-speaking, Mozabite, Russian
TXNRD3	15	Bantu-speaking, Biaka, Mozabite, Druze, Bedouin, Adygei, Basque, Orcadian, Russian, Balochi, Brahui, Hazara, Pathan, Yakut, Han	3	Mozabite, Adygei, Basque
SECISBP2	13	Bantu-speaking, Biaka, Yoruba, Mandenka, Druze, Sardinian, Russian, Sindhi, Balochi, Kalash, Yakut, Pima, Maya	3	Mandenka, Sardinian, Russian
AKR7L	6	French, Makrani, Balochi, Burusho, Naxi-Yi, Dai-Lahu	1	Burusho
GPx2	15	San, Bantu-speaking, Mandenka, Druze, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, Orcadian, Makrani, Balochi, Brahui, Burusho, Xibo-Mongolian, Japanese	4	BergamoItalian-Tuscan, Makrani, Balochi, Brahui
TXNRD2	18	San, Bantu-speaking, Biaka, Mandenka, Mozabite, Palestinian, Druze, Bedouin, French, Russian, Hazara, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, Han, Dai-Lahu, Papuan	1	Dai-Lahu
TRU-TCA2-1	1	Yoruba	0	
ARSB	16	Yoruba, Mandenka, Mozabite, French, Sindhi, Hazara, Burusho, Xibo-Mongolian, Oroqen-Hezhen-Daur, Japanese, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Papuan	5	Hazara, Burusho, Xibo-Mongolian, Oroqen-Hezhen-Daur, Japanese
SELENOM	10	San, Yoruba, BergamoItalian-Tuscan, Basque, Oroqen-Hezhen-Daur, JapaneseTu, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Pima	1	Yoruba

SELENOP	10	Bantu-speaking, Mbuti, Mandenka, Bedouin, BergamoItalian-Tuscan, Hazara, Pathan, Kalash, Uygur, Xibo-Mongolian	5	Bantu-speaking, Mandenka, Hazara, Pathan, Xibo-Mongolian
SELENOW	7	Yoruba, Bedouin, Brahui, Pathan, Xibo-Mongolian, Yakut, Papuan	0	
SEPHS2	4	Burusho, Xibo-Mongolian, NorthernHan-Tu, Maya	0	

4.4.1.4. Iron

In contrast to zinc, calcium and selenium-gene sets, the iron-associated gene set (n=44) shows signatures of positive selection that are less widespread amongst global regions or metapopulations. However, individual iron-associated genes show very strong evidence of positive selection that is somewhat shared over populations (**Table 4.7**). In particular, *ARHGEF3* and *FTMT* show strong signatures of positive selection according to both $Relate$ and F_{ST} in many Eurasian populations. This includes evidence of selection at the most stringent threshold for *FTMT* in the Yakut population of East Asia ($Relate\ pvalue = 3.37 \times 10^{-6}$) and for *HIF1A* in the Basque population of Europe ($Relate\ pvalue = 2.43 \times 10^{-6}$). It therefore appears that *ARHGEF3, FTMT* and *HIF1A* mediate iron-associated adaptation amongst different Eurasian populations. However, the significant and nearly-significant signatures of positive selection as inferred by $Relate$ are observed across Eurasia (**Table 4.8**), and hence these may not be strictly local adaptive responses, and may indeed be shared amongst Eurasian populations.

The strong signatures of positive selection identified at the *RHOA* gene appear to be more strongly indicative of a local response. These strong signatures of positive selection are identified in four East Asian populations, potentially consistent with an East-Asian response to an iron-associated selective pressure. Indeed, significant or nearly-significant signatures of positive selection (inferred by $Relate$) are only inferred in East Asian populations (**Table 4.8**). Hence, I suggest that the signatures of positive selection identified in *RHOA* represent an East-Asian specific adaptative response, potentially associated with iron levels.

There is additional evidence for some populations mediating iron-adaptation via an oligogenic response; the Biaka and Druze populations show a high number of signatures of positive selection as calculated by $Relate$, and the Bantu-speaking, Mandenka, Pima and Bougainville populations show a high number of signatures of positive selection as calculated by F_{ST} (**Table. S4.10-11**). The genes driving these signatures often differ from those identified as having strong evidence of selection (**Table 4.7**), and therefore these populations may mediate iron-associated pressures via small groups of different iron-associated genes. Hence, both a monogenic and oligogenic adaptive response to iron-associated selective pressures may be present amongst populations.

***Table 4.7**: **Iron-associated genes with p-values** $< \mathbf{10^{-5}}$. P-values calculated from the empirical distribution of either Relate or F_{ST}. P-values less than* 4.65×10^{-6} *(see **Section 3.4.5**) highlighted in bold.*

Gene	Population	*Relate* Significance	F_{ST} Significance
ARHGEF3	Bedouin		3.43e-5
	Palestinian		2.55e-5
	Basque	2.47e-5	
	Balochi		2.34e-5
	Brahui		6.51e-5
	Makrani		8.2e-5
	Dai-Lahu	6.80e-5	
FTMT	Adygei		7.54e-5
	Brahui	1.92e-5	
	Dai-Lahu		4.85e-5
	Yakut	**3.37e-6**	
RHOA	Han		8.62e-5
	Japanese		6.91e-5
	NorthernHan-Tu		9.99e-5
	She-Miao-Tujia		1.38e-5
TMPRSS6	Biaka	7.18e-5	
	Maya		4.9e-5
	Bougainville		8.23e-5
	Papuan		9.64e-5
HIF1A	**Basque**	**2.43e-6**	
	Sindhi	2.28e-5	
C19orf12	Adygei	2.01e-5	
CFAP251	Adygei	9.28e-5	
SLC40A1	Uygur	1.62e-5	
PLA2G6	Uygur	7.10e-5	
TAOK1	Bantu-speaking		3.42e-5
FTL	Mandenka		1.99e-5
HJV	Mandenka		2.91e-5
TFRC	Bougainville		3.80e-5
ACO1	Papuan		5.36e-5

***Table 4.8**: **The number ("No.") and name of the populations ("Populations") with signatures of positive selection**. Signatures of positive selection identified by the 1% or 0.1% tail of the empirical background distribution of Relate, for all iron-associated genes with p-values* $< 10^{-5}$ *in at least one population (**Table 4.7**)*

Gene	1% tail		0.1% tail	
	No.	Populations	No.	Populations
ARHGEF3	19	San, Bantu-speaking, Biaka, Mozabite, Bedouin, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, Makrani, Hazara, Kalash, Oroqen-Hezhen-Daur, Yakut, Han, NorthernHan-Tu, Dai-Lahu, Maya, Papuan	8	Bantu-speaking, Biaka, Adygei, Basque, Hazara, Oroqen-Hezhen-Daur, Dai-Lahu, Maya

FTMT	22	San, Mandenka, Druze, Adygei, Basque, French, Orcadian, Russian, Makrani, Sindhi, Balochi, Brahui, Pathan, Burusho, Uygur, Xibo-Mongolian, Yakut, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi, Pima	12	Mandenka, Druze, Adygei, Orcadian, Russian, Makrani, Balochi, Brahui, Xibo-Mongolian, Yakut, NorthernHan-Tu, She-Miao-Tujia
RHOA	3	Xibo-Mongolian, Han, NorthernHan-Tu	3	Xibo-Mongolian, Han, NorthernHan-Tu Papuan
TMPRSS6	14	San, Bantu-speaking, Biaka, Mozabite, Palestinian, Druze, Bedouin, Balochi, Brahui, Hazara, Burusho, Xibo-Mongolian, Surui-Karitiana, Papuan	2	Biaka, Druzee
HIF1A	25	Biaka, Yoruba, Mozabite, Palestinian, Adygei, BergamoItalian-Tuscan, Sardinian, Basque, French, Orcadian, Russian, Makrani, Sindhi, Brahui, Hazara, Pathan, Burusho, Kalash, Xibo-Mongolian, Japanese, NorthernHan-Tu, She-Miao-Tujia, Maya, Papuan, Bougainville	10	Biaka, Yoruba, Mozabite, BergamoItalian-Tuscan, Basque, French, Sindhi, Pathan, Burusho, Papuan
C19orf12	10	San, Biaka, Yoruba, Druze, Adygei, Basque, Orcadian, Sindhi, Brahui, Japanese	2	Adygei, Orcadian
CFAP251	11	Mbuti, Palestinian, Bedouin, Adygei, BergamoItalian-Tuscan, French, Russian, Kalash, Oroqen-Hezhen-Daur, NorthernHan-Tu, She-Miao-Tujia	2	Adygei, Kalash
SLC40A1	14	Bantu-speaking, Biaka, Palestinian, Druze, BergamoItalian-Tuscan, Basque, Makrani, Uygur, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Japanese, Han, She-Miao-Tujia	5	BergamoItalian-Tuscan, Uygur, Oroqen-Hezhen-Daur, Yakut, Japanese
PLA2G6	13	San, Mbuti, Biaka, Palestinian, Druze, BergamoItalian-Tuscan, Makrani, Balochi, Pathan, Kalash, Uygur, Yakut, NorthernHan-Tu	3	Mbuti, Uygur, Yakut
TAOK1	6	Bantu-speaking, Biaka, Hazara, Han, She-Miao-Tujia, Dai-Lahu	1	Han
FTL	2	Mozabite, Dai-Lahu	0	
HJV	0		0	
TFRC	7	Mbuti, Yoruba, Oroqen-Hezhen-Daur, Yakut, NorthernHan-Tu, She-Miao-Tujia, Surui-Karitiana	2	NorthernHan-Tu, She-Miao-Tujia
ACO1	12	San, Biaka, Yoruba, Mandenka, Mozabite, Makrani, Brahui, Pathan, Xibo-Mongolian, Oroqen-Hezhen-Daur, Yakut, Han	2	Biaka, Yakut

4.4.1.5. Iodine

In comparison to all other zinc, calcium, selenium and iron-associated gene sets, there are more limited signatures of positive selection within the iodine-associated gene set (n=18), particularly when isolating the strong signatures of positive selection (**Table 4.9**). These signatures of positive selection are also considerably less widespread compared to those of the previous micronutrient gene sets, but are still observed amongst some isolated African, European, Middle-Eastern and Central-South Asian populations (**Table 4.9**). *THRB* shows the strongest signature of positive selection (bypassing the most stringent threshold; $Relate\ pvalue = 3.23 \times 10^{-6}$) in the Palestinian population of the Middle-East, but this population does not appear to show evidence for positive selection at the gene set level (**Tables S4.12-13**). Given this strong signature in the Palestinian population and the number of nearly-significant signatures of positive selection (**Table 4.10**), *THRB* is the strongest candidate gene for iodine-associated adaptation.

Still, the more geographically concentrated signatures of positive selection, and the geographic patterns of signatures of positive selection across all populations (**Fig. 4.1-2**), hence suggest that iodine-associated adaptation is more localised in comparison to zinc, calcium, selenium and iron.

The Maya population of the Americas and Uygur population of Central-South Asia have the strongest evidence of iodine-associated selection at the gene set level (**Fig. 4.1-2**). Five and four genes are identified with signatures of positive selection in the Maya according to *Relate* and F_{ST}, respectively, and five genes are identified with signatures of positive selection in the Uygur according to F_{ST}. However, for the latter population, this includes signatures of positive selection identified in the *DIO1* and *DIO2* genes, which are also associated with selenium metabolism, and therefore these signatures may instead capture selenium-associated adaptation. There are also no iodine-associated genes identified with signatures of positive selection in the Uygur population according to *Relate*. Hence, the Maya populations presents the strongest evidence for adaptation in response to iodine levels.

Further, of the iodine-associated genes exhibiting signatures of positive selection in the Maya (as inferred using *Relate*), four also show signatures of positive selection in the Mbuti population (**Tables S4.12-13**). Three of these genes are thyroid receptors (*THRA, THRB, TRIP4*) and are associated with both iodine metabolism and growth pathways. Given the short stature of these populations and strong signatures of positive selection shared on genes known to affect height, this provides some support for the link between iodine-associated adaptation and short stature (as suggested in (Herráez et al. 2009)).

***Table 4.9**: **Iodine-associated genes with p-values** $< \mathbf{10^{-5}}$**.** P-values calculated from the empirical distribution of either Relate or F_{ST}. P-values less than 4.65×10^{-6} (see **Section 3.4.5**) highlighted in bold.*

Gene	Population	*Relate* Significance	F_{ST} Significance
TSHR	Mandenka		8.17e-5
	Palestinian		7.98e-5
THRB	Sardinian	2.94e-5	
	Palestinian	**3.23e-6**	

TRIP4	Mbuti		3.96e-5
TPO	Mozabite	2.5e-5	
SLCO1C1	BergamoItalian-Tuscan	7.78e-5	
SLC5A5	Orcadian	9.83e-5	
SECISBP2	Sardinian	3.27e-5	
SLC16A2	Brahui	1.51e-5	

***Table 4.10**: **The number ("No.") and name of the populations ("Populations") with signatures of positive selection**. Signatures of positive selection identified by the 5% or 0.1% tail of the empirical background distribution of Relate, for all iodine-associated genes with p-values $< 10^{-5}$ in at least one population (**Table 4.9**)*

Gene	1% tail		0.1% tail	
	No.	Populations	No.	Populations
TSHR	17	Bantu-speaking, Biaka, Yoruba, Mandenka, Palestinian, Druze, Sardinian, Basque, French, Orcadian, Russian, Balochi, Pathan, Burusho, Uygur, Maya, Papuan	2	Basque, Maya
THRB	27	San, Bantu-speaking, Biaka, Mozabite, Palestinian, Druze, BergamoItalian-Tuscan, Sardinian, Basque, French, Orcadian, Russian, Sindhi, Brahui, Hazara, Pathan, Burusho, Kalash, Xibo-Mongolian, Yakut, Japanese, She-Miao-Tujia, Naxi-Yi, Dai-Lahu, Pima, Maya, Surui-Karitiana	12	Bantu-speaking, Palestinian, Sardinian, Orcadian, Sindhi, Burusho, Kalash, Xibo-Mongolian, Naxi-Yi, Dai-Lahu, Maya, Surui-Karitiana
TRIP4	15	San, Bantu-speaking, Yoruba, Mozabite, Palestinian, Druze, BergamoItalian-Tuscan, Russian, Makrani, Japanese, Han, Tu, NorthernHan-Tu, Dai-Lahu, Pima, Maya	3	Mozabite, Han, Maya
TPO	4	Mandenka, Mozabite, Bedouin, Makrani	1	Mozabite
SLCO1C1	20	Mbuti, Biaka, Yoruba, Mandenka, Mozabite, Palestinian, Druze, BergamoItalian-Tuscan, Basque, Sindhi, Balochi, Brahui, Hazara, Pathan, Burusho, Oroqen-Hezhen-Daur, Yakut, Japanese, Han, NorthernHan-Tu	11	Biaka, Palestinian, BergamoItalian-Tuscan, Basque, Sindhi, Balochi, Brahui, Hazara, Burusho, Oroqen-Hezhen-Daur, Han
SLC5A5	10	Bantu-speaking, Biaka, Yoruba, Mozabite, French, Orcadian, Pathan, Kalash, NorthernHan-Tu, Naxi-Yi	1	Orcadian
SECISBP2	13	Bantu-speaking, Biaka, Yoruba, Mandenka, Druze, Sardinian, Russian, Sindhi, Balochi, Kalash, Yakut, Pima, Maya	3	Mandenka, Sardinian, Russian
SLC16A2	9	Bantu-speaking, Mandenka, Mozabite, Brahui, Yakut, Han, NorthernHan-Tu, She-Miao-Tujia, Naxi-Yi	1	Brahui

4.4.2. Co-Occurring Signatures of Positive Selection

If genes are functionally linked and frequently co-demonstrate signatures of positive selection within populations, this may indicate groups of genes responding to the same selective pressure in different human groups (Berg and Coop 2014; Berg, Zhang, et al. 2019; Lewis et al. 2020). Hence, to identify potential pathways for micronutrient response, or which groups of genes may be co-adapting, networks are built representing genes that share signatures of positive selection (as inferred using *Relate*) over the same populations (**Fig 4.3**). Observations from these networks are summarised below.

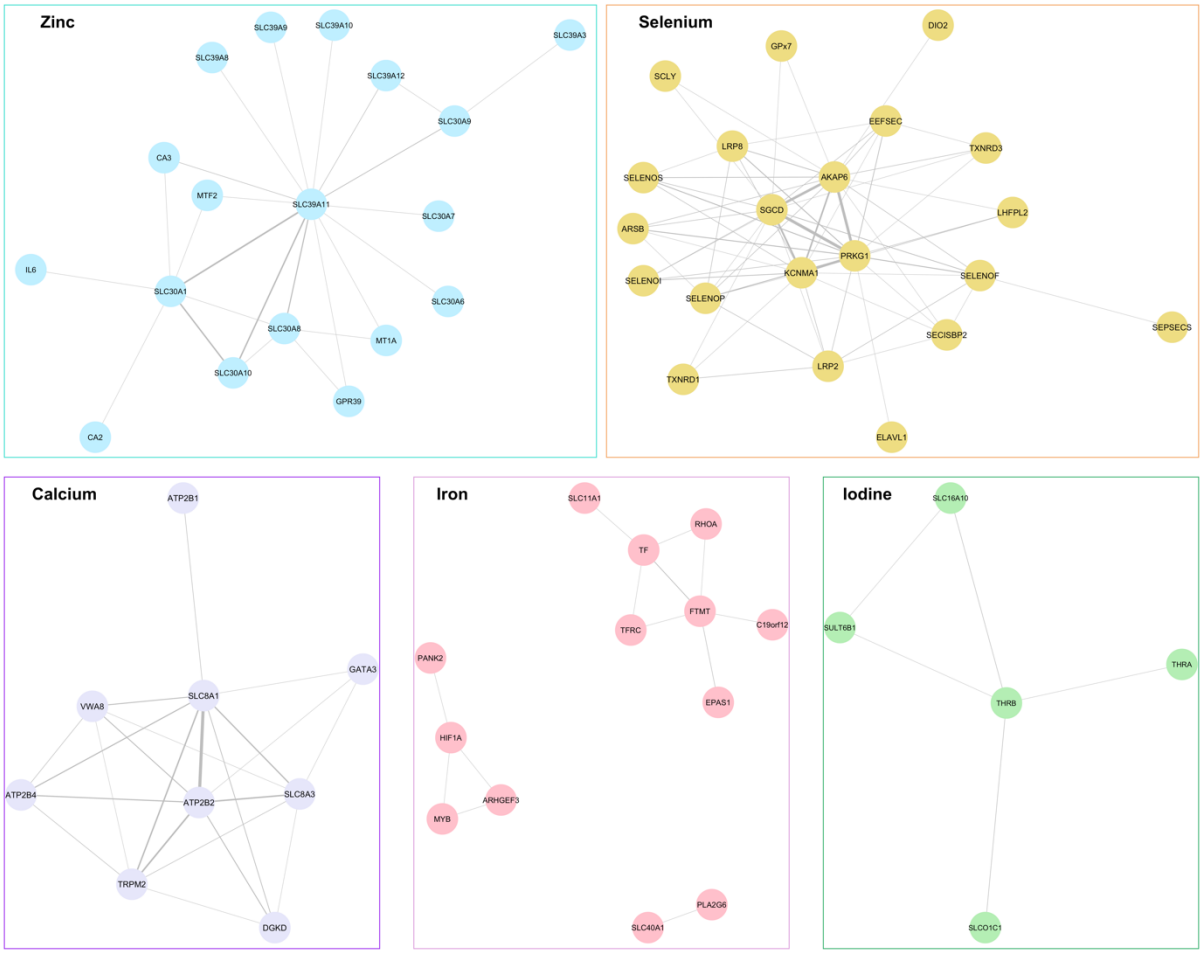


***Figure 4.3**: **Gene networks for zinc, selenium, calcium, iron and iodine-associated genes.** Genes are connected if they share signatures of positive selection as identified using Relate in two or more populations; thickness of the connecting lines corresponds to the number of populations where signatures of positive selection are shared.*

The zinc-associated gene network recapitulates **Section 4.4.1.1**, emphasising the prevalence of signatures of positive selection in zinc-transporter genes in the *SLC30* and *SLC39* families and the likelihood of multiple zinc-associated genes mediating an adaptive response. I observe the central role of *SLC39A11,* and identify frequently co-occurring signatures of positive selection between this gene and *SLC30A8*, *SLC30A10* and *SLC30A1.* The co-occurring signatures of positive selection amongst these genes are identified amongst African, European, Central-South Asian and East Asian populations, and hence this network of genes may mediate widespread zinc-associated adaptation. Given that there are also less frequent co-occurring signatures of positive selection amongst other zinc-transporter genes (**Table 4.2**), it is possible that some zinc-transporter genes might be interchangeable in their ability to mediate adaptation or that the genomic nature of adaptation in response to zinc levels is diverse over populations. Still, as outlined in **Section 4.4.1.1**, I suggest that many zinc transporter genes are involved in mediating the adaptive response to zinc levels, but some zinc-transporter genes have stronger evidence or a more strongly supported role in such adaptation.

In the calcium and selenium gene sets, some smaller groups of genes show frequently co-occurring signatures of positive selection. In the calcium gene set, *ATP2B2* and *SLC8A1* particularly share signatures of positive selection in the same populations, alongside less frequent co-occurring signatures between these genes and *SLC8A3* and *TRPM2*.

The selenium gene set also appears to show a central network of genes co-exhibiting signatures of positive selection in the same populations (*SGCD, AKAP6, PRKG1* and *KCNMA1*), which thus may largely mediate adaptation in response to selenium levels. These co-occurring signatures are observed in close populations and populations from very different regions (*e.g.,* they co-occur in multiple African, Middle-Eastern, European, Central-South Asian and East Asian populations), hence appearing to be a gene set globally mediating adaptation. However, there are many other co-occurring signatures of positive selection amongst other selenium-associated genes (often shared in multiple African, Central-South Asian and East Asian populations; see **Table 4.6**). Whilst there are a large number of selenium-associated genes in this study, these observations are also in accordance with the suggested oligogenic or polygenic nature of selenium-associated adaptation (White et al. 2015): here, adaptation may be primarily mediated by allele frequency changes in a small network of genes exhibiting strong evidence of selection (**Section 4.4.1.3**), accompanied by more moderate allele frequency changes in additional, perhaps more constrained, selenium-associated genes.

The iron gene network partitions into three clusters, with the top candidate genes (*FTMT, RHOA, HIF1A* and *ARHGEF3*) split over two of these clusters. This demonstrates that signatures of positive selection do not often co-occur between the same genes across different populations, and it appears that the genes which mediate iron-adaptation may differ across different populations, as discussed in **Section 4.4.1.4**. The connection of *FTMT* and *RHOA* in the network, due to shared signatures of positive selection in only East Asian populations, is a product of *RHOA* only exhibiting signatures of positive selection in East Asian populations.

Finally, there are no frequently co-occurring signatures of positive selection amongst iodine-associated genes, further supporting isolated pockets of adaptation via different genes as suggested in **Section 4.4.1.5.**

4.4.3. Geographically Global Patterns of Adaptation

Some ZCSII-associated genes, particularly zinc-associated genes, exhibit a high degree of differentiation from Yoruba (as calculated from F_{ST}) in many non-African populations (**Section 4.4.1**), which is interpreted as a shared signature of positive selection. I suggest that this is most likely due to positive selection on a common non-African ancestral population (such as a migrating Out of Africa population), rather than positive selection in many non-African populations. Other ZCSII-associated genes show signatures of positive selection that are concentrated at the metapopulation level, *e.g.,* the selenium-associated signatures shared amongst populations in East-Asia, which may also be due to a shared selective pressure or positive selection having acted on a common ancestral population of East Asians.

To investigate this further, haplotype networks of identified genes of interests are built, partitioning haplotypes by metapopulation. This helps us to visualise the genetic variation across populations, and infer if putative selection was likely on the same (or very similar haplotypes) or on very different haplotypes. Thus, most explicitly, these haplotype networks distinguish between putative selection on *de novo* mutation (the same haplotype background) and putative selection on standing variation (various haplotype backgrounds). However, if different metapopulations show uniform but divergent haplotypes, this suggests convergent selection between these metapopulations, rather than a shared selection event in the common ancestor of these metapopulations. These haplotype networks can also be used to identify potential recombination amongst haplotypes, which will be represented by cycles in the network.

Genes of interest are identified as those with signatures of positive selection, particularly identified by F_{ST}, in the most populations (**Fig. S4.1**), and hypothesise that these genes may have undergone adaptation in an ancestral non-African population. Haplotypes of length 10kb and 20kb were built around a focal SNP (**Table S4.2**) which are chosen as described in **Section 4.3.4.**

4.4.3.1. Adaptation Out of Africa

The zinc-associated genes showing strong differentiation with respect to Yoruba in many populations (*SLC39A4*, *GPR39*, *SLC30A9, SLC39A11* and *SLC39A14)* all show more diverse haplotypes in African populations (red in **Fig 4.4-5; Figs. S4.2-11**) compared to non-African populations (as expected from increased genetic diversity in Africans (Campbell and Tishkoff 2008; Tucci and Akey 2019)). The focal SNPs of *SLC30A9 and SLC39A11* are found in identical, or highly similar, haplotypes at high-frequency in non-African populations, in line with expectations under positive selection increasing the frequency of a beneficial allele in an ancestral non-African population (**Fig 4.4**, where the exact haplotypes carrying the focal, putatively selected variant for *SLC30A9* is shown in **Fig 4.5**). Here, I suggest that selection is most likely from a low-frequency allele (either from a *de novo* mutation or an allele segregating at low frequency). The *ATP2B2* and *ATP2B4* genes both show identical or highly similar haplotypes shared amongst the majority of non-Africans (**Fig 4.6; Figs. S4.12-17**). The cluster of closely related, similar haplotypes (particularly observed for *ATP2B4*) indicate that selection may have acted on more varied genetic backgrounds and therefore more suggestive of selection acting on standing variation.

However, the remaining focal SNPs of zinc-associated genes are found in multiple clusters of identical or highly related haplotypes shared amongst non-Africans but of which are highly divergent from each other (see **Fig 4.4, Figs. S4.4.2-11**). This divergence of common haplotypes, partnered with no clear sorting amongst metapopulations, suggests that if selection was indeed present, it most likely acted on different genetic backgrounds (*i.e.,* selection on standing variation, where segregating SNPs were likely at appreciable frequencies), perhaps in an ancestral non-African population.

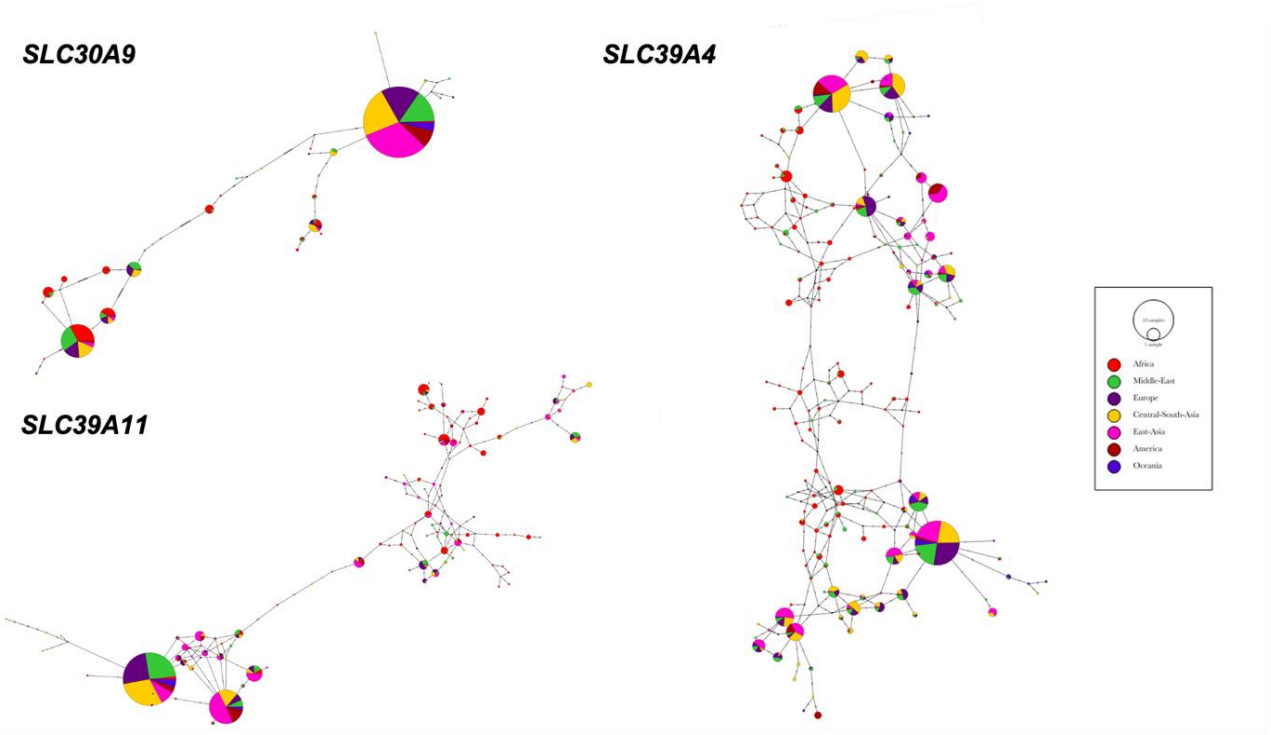


Fig 4.4: Haplotype networks built from the 20kb region surrounding focal SNPs of zinc-associated genes. *Shown for SLC30A9 (position: 42004040), SLC39A4 (position: 144414297) and 10kb surrounding the focal SNP of the zinc-associated gene SLC39A11 (position: 73010373).*

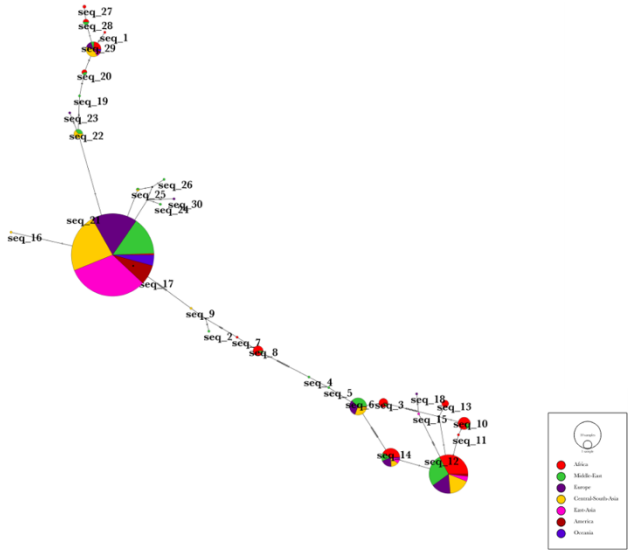


Fig 4.5: Haplotype networks with labelled sequences built from the 20kb region surrounding focal SNP (position: 73010373) of SLC39A11. *Sequences containing the focal SNP with the putatively selected variant are seq_16, seq_17, seq_21, seq_22, seq_23, seq_24, seq_25, seq_26, seq_30. Gene chosen for its relative visual clarity when viewing sequence numbers.*

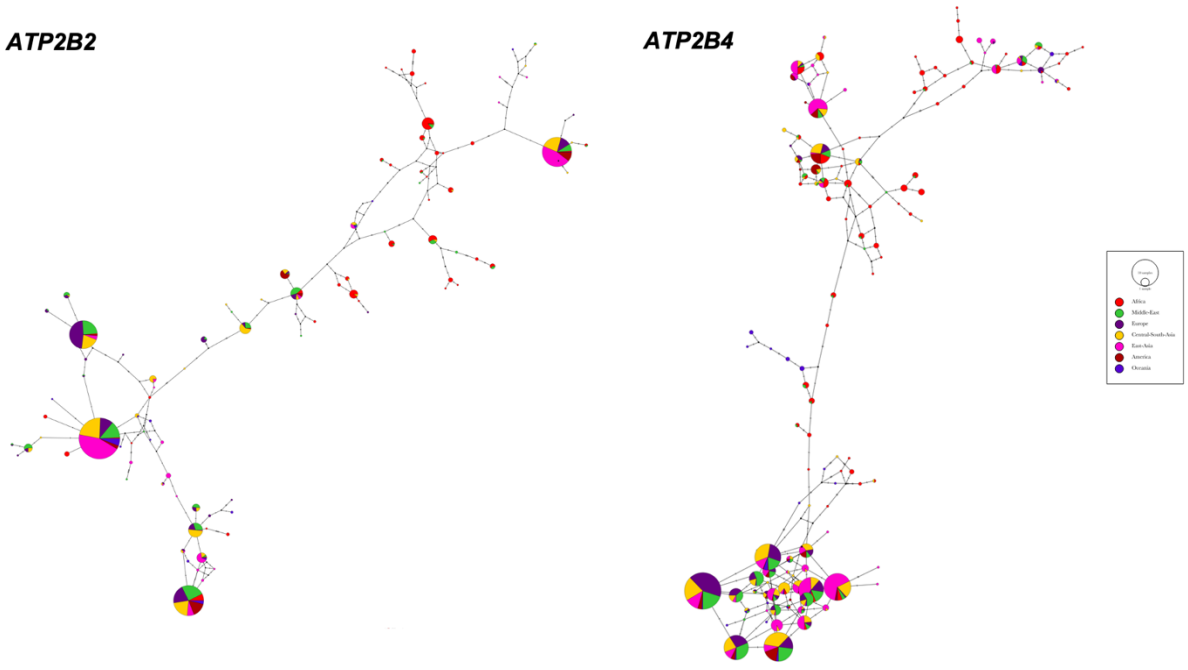


Fig 4.6: Haplotype networks built from the regions surrounding focal SNPs of calcium-associated genes*. Shown for ATP2B2 (position: 10636328; 20kb region) and 10kb ATP2B4 (position: 203667951; 10kb region)*

4.4.3.2. Adaptation within Metapopulations

The focal SNPs of the selenium-associated genes *PRKG1, EEFSEC* and *AKAP6* are within haplotypes which appear to cluster, at least in some degree, by metapopulation. Specifically, these SNPs are found within identical or highly similar haplotypes in particularly East Asian and African individuals (pink in **Fig 4.7, Figs S4.18-26**), but within otherwise variable haplotypes in other metapopulations (**Fig 4.7, Figs. S4.18-26)**. Most strikingly, the selenium-associated gene *SGCD* shows an identical haplotype carrying the focal and putatively selected variant at high frequency in the East Asian metapopulation ("seq 18" as labelled in **Fig 4.8**).

This pattern of genetic variation (*i.e.,* uniform haplotype structure within individual metapopulations) is as would be expected for selection acting convergently on these genes in East Asian and African populations, rather than shared amongst all populations. The divergence of some haplotypes of high frequency amongst either East Asian or African individuals also suggests that the selected allele was present in multiple haplotypes when selection started, suggesting SSV and an appreciable frequency of the selected allele.

The haplotypes containing the focal SNPs of the iron-associated gene *ARHGEF3* also appear to group by metapopulation (particularly in East Asian populations), in support of selection focused in Eurasia and not a result of a selection event in an ancestral non-African population. Given that the haplotypes are more diverse, I suggest that selection, if present, was selection on standing variation within these populations (**Fig S4.27-32**).

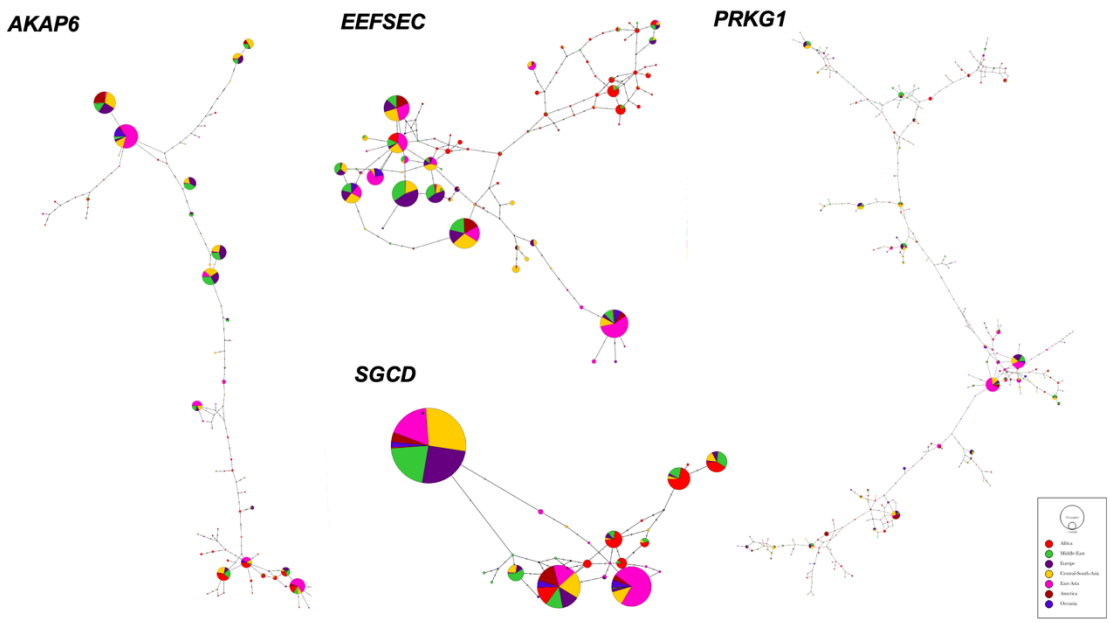


Fig 4.7: Haplotype networks built from the 20kb region surrounding focal SNPs of selenium-associated genes*. Shown for EEFSEC (position: 128412869), SGCD (position: 156057959), PRKG1 (position: 51471686) and AKAP6 (position: 32446036)*

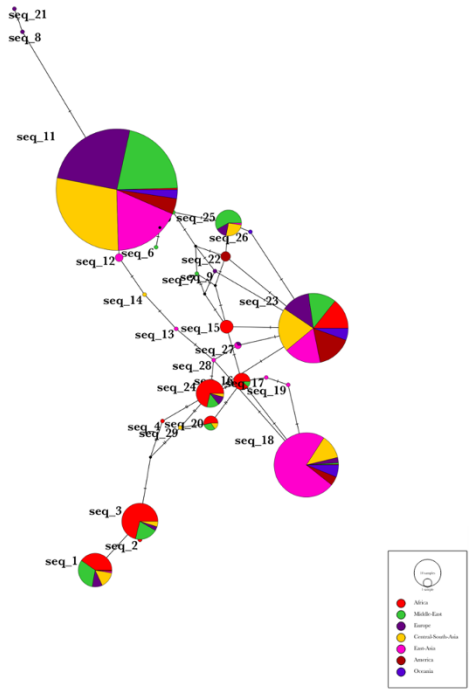


Fig 4.8: Haplotype networks with labelled sequences built from the 20kb region surrounding focal SNP (position: 156057959) of SGCD. *Sequences containing the focal SNP with the putatively selected variant are seq_13, seq_15, seq_16, seq_18, seq_20, seq_22, seq_23, seq_24, seq_27, seq_28, seq_29. Gene chosen for its relative visual clarity when viewing sequence numbers.*

4.4.4. Estimating the Onset of Selection

I now question whether signatures of positive selection identified on candidate genes were more likely driven by selective pressures exerted when encountering new environments, or as a result of more recent cultural changes. This analysis is heavily computationally intensive, so it was not possible to run it for all genes and micronutrients. I focus on calcium (*ATP2B2, ATP2B4, SLC8A1, SLC8A2* and *SLC8A3*) and iron-associated genes (*FTMT, ARHGEF3*, *HIF1A* and *SLC40A1*) with the strongest evidence of selection, since levels of these micronutrients have suggested to have been particularly affected by the transition to the Neolithic diet (Dobrovolskaya 2005; Naugler 2008; Gerbault et al. 2011, 2011).

To answer this question, I first identify new focal SNPs (given in **Table S4.3**) which are chosen as described in **Section 4.3.4.** *CLUES* is then used (Stern et al. 2019) to infer the log-likelihood ratio of selection acting at one of four different timepoints (14kya, 28kya, 42kya and 56kya: timepoints that encompass the time just following the Out of Africa migration, the time of migrations into new Eurasian environments, and the time just preceding the Neolithic transition) (**Figs. S4.33-51**). I then verify the onset of selection suggested by this programme by the allele trajectories through time as reconstructed using *Relate* (Speidel et al. 2019).

4.4.4.1. Onset of Calcium-Associated Selection

The evidence of positive selection inferred from *CLUES* largely agrees with previously described evidence from *Relate* and F_{ST}; evidence for positive selection (log-likelihood ratios of selection > 4) is inferred across all focal SNPs of the candidate calcium-associated genes and in many populations with previously identified signatures of positive selection (**Table 4.11**). For some genes, novel signatures of positive selection are identified in some populations, *i.e.,* those not identified using *Relate* or F_{ST} (indicated in **Table 4.11**).

The highest log-likelihood ratios of selection are observed in the Middle-Eastern Mozabite (*ATP2B2* and *ATP2B4*) and Bedouin (*SLC8A1* and *SLC8A2*) populations, with estimated selection coefficients of ~0.003. The strongest selection coefficients (also accompanied by log-likelihood ratios indicative of selection) across the entire set of calcium focal SNPs and populations are observed in the Central-South Asian Makrani (*ATP2B2,* s~0.0087) and the European Orcadian (*SLC8A3,* s~0.0057) populations. This latter population does not have a previously identified signature of positive selection in this gene according to *Relate* or F_{ST}, but does have a nearly-significant signatures of positive selection according to *Relate* (**Table 4.4**). Hence, this provides additional support for these populations undergoing calcium-associated adaptation.

If only including the focal SNPs and populations for which *CLUES* suggests selection (log-likelihood ratios of selection > 4), the log-likelihood ratio of positive selection generally increases further back in time. However, there is often very little difference between inferred log-likelihood ratios of selection at the 42kya and 56kya timepoints (**Fig. 4.9, Figs. S4.33-42**), and so the exact onset of proposed selection cannot be confidently proposed. Moreover, the evidence of selection is compared at relatively close timepoints with limited data for each locus, and therefore only broad patterns can be examined here. These inferences are considered with the inferred allele frequency trajectories calculated using *Relate* to suggest that the onset of selection is more likely around 40 – 30kya in the majority of populations with significant signatures of positive selection, suggesting that selection could have accompanied the colonising of new Eurasian environments.

The allele frequency increase of the focal SNPs is, in most populations, inferred to have begun earlier than 10kya. But exceptions are observed in *ATP2B4 and SLC8A3*. The Middle-Eastern populations, particularly the Mozabite, show additional sharp increases of frequency of the focal SNPs of *ATP2B4* (positions: 20364823, 203667951) around 10kya. There are sharp increases of the frequency of a focal SNP of *SLC8A3* (position: 70182346) in the Mozabite and Pima populations between 10kya and 5kya. These results suggest that besides widespread positive selection before 10kya, additional later calcium-associated adaptations may have occurred in particular (particularly the Mozabite) populations, possibly due to major dietary changes.

Table 4.11: Populations with log-likelihood ratios of selection > 4 for calcium-associated genes of interest. *Calculated for given times of the onset of selection ("Time") and shown alongside their inferred selection coefficients, for focal SNPs of calcium-associated genes of interest. Populations marked with * do not have previously identified signatures of selection (identified by the 0.1% tail of the empirical distribution of either Relate or F_{ST}).*

Gene	Position	Population	Time (kya)	Log Likelihood Ratios	Selection Coefficient
ATP2B2	10604833	Mozabite	56	6.8143	0.00288
			42	6.7209	0.00293
			28	6.0841	0.00308
			14	4.7436	0.00386
		Sardinian	54	6.1902	0.003
			42	6.1464	0.0032
			28	5.5478	0.00305
		Makrani	56	5.8721	0.00864
			42	5.8704	0.00864
			28	5.8537	0.00869
			14	5.7553	0.00981
		Bedouin	56	5.1878	0.00222
			42	5.1107	0.00228
			28	4.4327	0.00237
		Basque	56	4.7394	0.00659
			42	4.7382	0.00059
			28	4.7232	0.00659
			14	4.3041	0.00671
		Palestinian	56	4.1905	0.00203
			42	4.0711	0.00205
ATP2B4	20364823	Mozabite	28	5.1804	0.0038
			42	5.1205	0.00298
			56	4.6161	0.00278
			14	4.158	0.00601
		Mandenka*	56	4.585	0.00219
			42	4.5796	0.00228
			28	4.2572	0.00269
		Druze	42	4.1131	0.00237
			56	4.1117	0.00221
		Yoruba*	56	4.0166	0.00166
	203667951	Druze	56	5.483	0.00226
			42	5.0879	0.00228
			28	4.1476	0.00247
		Bedouin	56	5.4379	0.00211
			42	5.1408	0.00232
			28	4.0238	0.0025
		Sindhi	56	4.6668	0.0019
SLC8A1	40584510	Bedouin	56	7.5284	0.00226
			42	7.0133	0.00225
			28	5.752	0.00242
			14	5.0701	0.00337
		Basque*	56	4.3386	0.00254
			42	4.2241	0.00269
		Makrani	56	4.3009	0.00226

			42	4.0505	0.00234
SLC8A1	40394610	Hazara	56	5.2674	0.00322
			42	5.2519	0.00327
			28	5.0758	0.0041
		Yakut	56	4.5985	0.00242
			42	4.3783	0.00254
SLC8A2	47428756	Bedouin	56	6.3259	0.00234
			42	6.2555	0.00238
			28	5.1166	0.00283
	47437107	Brahui	56	4.8037	0.00166
			42	4.3034	0.00173
SLC8A3	70175561	Orcadian*	56	5.8316	0.00576
			42	5.8307	0.00576
			28	5.8133	0.00574
			14	4.7667	0.00596
		French*	56	4.5841	0.00238
			42	4.488	0.00247
		NorthernHan-Tu	56	4.3755	0.00284

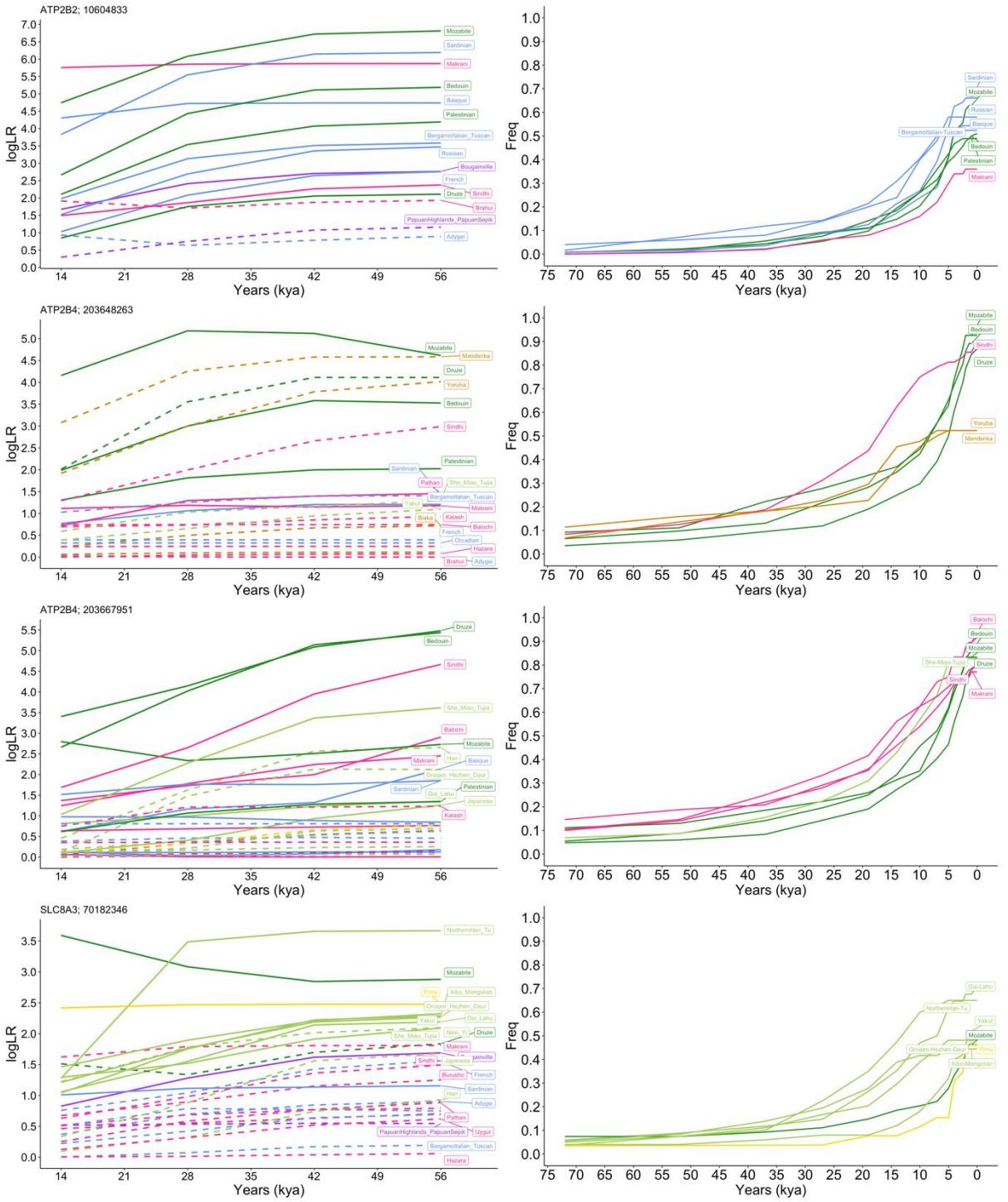


Fig. 4.9: Inferred likelihood ratios of selection and allele frequency over time for calcium-associated genes of interest. *Left: Inferred log likelihood ratios for ATP2B2, ATP2B4 and SLC8A3 focal SNPs for populations with pvalues < 0.05, as calculated according to the empirical distributions of either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Right: Inferred frequency trajectory of the same focal SNP for populations of interest (other populations omitted for clarity). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

4.4.4.2. Onset of Iron-Associated Selection

The evidence of positive selection inferred from $CLUES$ (log-likelihood ratios of selection > 4) again largely supports previously described evidence from $Relate$ and F_{ST} (**Table 4.12**). The strongest evidence of positive selection inferred from $CLUES$ is for the focal SNP of *HIF1A* (position: 61709502) in the Palestinian population, where the log-likelihood ratios for selections is > 7 for selection acting at 28ky, 42kya and 56kya (the highest log-likelihood ratio value and is therefore strongest evidence of selection as calculated by $CLUES$ over all calcium and iron-associated SNPs; **Table 4.12**). The strongest evidence of positive selection, as inferred by $CLUES$, for other focal SNPs include *FTMT* in the Brahui (position: 121846819; log-likelihood ratio= 6.50) and *HIF1A* in the Basque (position: 61741756; log-likelihood ratio= 5.09, distinct from other populations; see **Fig. 4.10**).

Again, very small differences between log-likelihood ratios of selection calculated for the timepoints 28kya, 42kya and 56kya are observed across most populations. Populations of interest, inferred as such from **Table 4.12** and their allele frequency trajectories, do show a general increase in frequency of the focal SNPs around 20kya – 30kya (**Fig 4.8; Figs. S4.43-51**)**.** This, and the uniformity of log-likelihood ratios across 28-52kya, could suggest a slightly later onset of selection in iron-associated genes compared to that inferred for the majority of calcium-associated genes (estimated as 30-40kya in the majority of populations), perhaps as a result of more recent and smaller scale migrations into Eurasian environments. This agrees with the Eurasian-specific signatures of positive selection in iron-associated genes, see **Section 4.4.1.4**.

On the other hand, allele frequency increases are inferred to be very recent for the focal SNPs of some populations. The focal SNP of *HIF1A* (position: 61741756) in the Basque population is inferred to increase rapidly in frequency between 10kya and 5kya, as does the *FTMT* focal SNP (position: 121846819) in the Mozabite population. The log-likelihood ratios observed in the focal SNP of the *HIF1A* gene in the Basque population suggest that this population might have undergone recent iron-associated adaptation. Finally, very sharp increases of the focal SNP of *ARHGEF3* (position: 57043874) are observed in the Maya and Dai-Lahu population at 20kya. Whilst log-likelihood ratios are under 4 (given as 2.7194-3.922 in the Maya across the four timepoints, 2.1397-2.4992 in the Dai-Lahu), this increase is at the level of which could suggest selection acting at this time on the ancestors of these populations.

Table 4.12: Populations with log-likelihood ratios of selection > 4 of iron-associated genes of interest. *Calculated for given times of the onset of selection ("Time") and shown alongside their inferred selection coefficients, for given focal SNPs of iron-associated genes of interest. Populations marked with * do not have previously identified signatures of selection (identified by the 0.1% tail of the empirical distribution of either Relate or F_{ST}).*

Gene	Position	Population	Time (kya)	Log Likelihood Ratio	Selection Coefficient
FTMT	121846819	Brahui	56	6.4986	0.0024
			42	6.063	0.00245
			28	5.1155	0.00273
			14	4.0101	00042
		Druze	56	5.4083	0.0021
			42	4.8906	0.00206
		Yakut	56	5.1566	0.00286
			42	4.9626	0.00303
		Hazara*	56	4.235	0.00242
			42	4.1336	0.00249
HIF1A	61687412	She-Miao-Tujia	56	4.2687	0.00256
	6170952	Palestinian	56	8.7594	0.00302
			42	8.7443	0.00303
			28	7.8056	0.00302
			14	5.0698	0.00352
		Sindhi	56	5.0616	0.0019
			42	4.5131	0.00187
		Pathan	56	4.403	0.00198
	61741756	Basque	56	5.0855	0.00314
			42	4.9233	0.00315
			14	4.0508	0.00586
			28	4.0096	0.00354

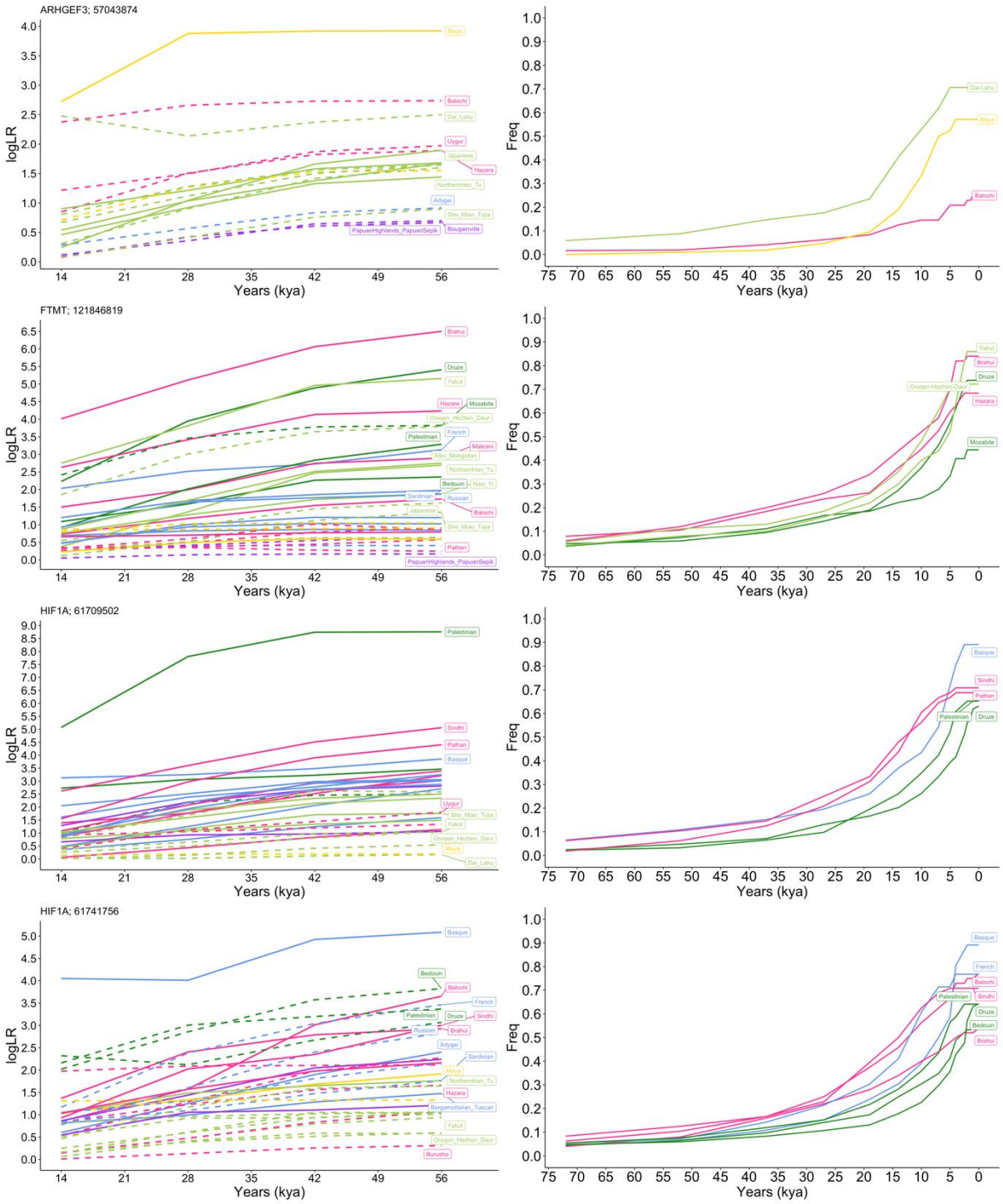


Fig. 4.10: Inferred likelihood ratios of selection and allele frequency over time for iron-associated genes of interest. *Left: Inferred log likelihood ratios for ARHGEF3, FTMT, HIF1A and SLC40A1 focal SNPs for populations with pvalues < 0.05, as calculated according to the empirical distributions of either either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Right: Inferred frequency trajectory of the same focal SNP for populations of interest (other populations omitted for clarity). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

4.5. Discussion

Micronutrient levels in the diet have been inferred as likely drivers of selection across modern human populations (Engelken et al. 2014, 2016; Sverrisdóttir et al. 2014; White et al. 2015; Ye et al. 2015; Zhang et al. 2015a), a hypothesis that is supported by the work presented in **Chapter 3**. In that chapter, I propose that micronutrient-associated adaptation has contributed to modern human genetic diversity (**Chapter 3**) and outlined which micronutrients had the strongest evidence for acting as a selective driver during the history of our species.

In this chapter, the previous evidence is re-visited to discuss five micronutrients, investigating the associated signatures of positive selection and the evolutionary history of these genes in further detail. Deficient levels of these five micronutrients, four trace minerals (zinc, selenium, iron, iodine) and one macromineral (calcium), result in a series of severe health issues, of which are common across modern human populations and perhaps also within our evolutionary history (Kelly and Snedden 1960; Xia et al. 2005; Manning et al. 2012; Bailey et al. 2015; Khan et al. 2022; Shlisky et al. 2022). These micronutrients are not only likely selective drivers, but their associated genes present good evidence for having undergone adaptation in modern human populations (**Chapter 3**), and are hence good candidates to further explore. I thus ask which genes are most likely to mediate micronutrient-associated adaptation and in which populations, how adaptation of different genes may co-occur, and the most likely timing of such adaptation (extrapolating to infer the most likely selective drivers).

I show that the inferred signatures of positive selection are often the strongest in only a few genes of a given micronutrient-associated gene set, building on the work undertaken in **Chapter 3** suggesting that micronutrient-associated adaptation is likely oligogenic in nature. In some cases, the candidate genes showing the strongest evidence of positive selection have signatures shared over global geographic regions, as expected if selection occurred on populations ancestral to these extant populations. In other micronutrient-associated genes, particularly those associated with iron and iodine, signatures are more geographically restricted as if positive selection occurred in local pockets across the globe. Finally, from the inferred allele frequency trajectories and likelihood of selection onset over time, the most likely timing of adaptation associated with two micronutrients (iron and calcium) is inferred. The specifics of these inferences for each micronutrient are summarised below.

4.5.1. Zinc

Signatures of positive selection have been identified in multiple zinc-transporter genes in many previous studies (*e.g.,* (Engelken et al. 2014; Zhang et al. 2015a; Roca-Umbert et al. 2022)). Much of this literature has highlighted *SLC30A9* and *SLC39A4* as those with the strongest evidence of selection (Engelken et al. 2014; Zhang et al. 2015a), which is recapitulated here. Both *SLC39A4* and *SLC30A9* show strong signatures of positive selection (as calculated by F_{ST}) across many populations, where the evidence of selection is strongest in the Makrani of Central South Asia (*SLC39A4*) and Han of East Asia (*SLC30A9*). The Makrani population live in modern-day Pakistan, where zinc deficiency is prevalent (22.1%) and up to 96.1% of grain samples are zinc-deficient (Rehman et al. 2020; Ishfaq et al. 2021). Equally, zinc levels are low in the calcareous soil of China (Karim et al. 2012), and approximately 100 million people are affected by zinc deficiency in this

region (Ma et al. 2008). These recorded deficiencies are in agreement with these populations facing strong selective pressure from zinc levels.

I identify ten further zinc-transporter genes showing signatures of positive selection that are identified amongst all metapopulations excluding Oceania: *SLC39A4, SLC30A9, SLC39A11, SLC39A12, SLC30A7, SLC30A8, SLC30A1, SLC39A10, SLC39A14* and *SLC30A10* (randomly distributed across the genome, hence these are not shared signatures). Again, strong signatures of positive selection are observed in the Makrani population, particularly in *SLC39A11.*

However, these shared signatures of positive selection do not necessarily represent the independent adaptive responses of individual populations to varied zinc content in soils across the globe. Rather, the sharing of signatures of positive selection across populations most likely reflects a shared selective pressure experienced by many populations (or by a common ancestor of many populations). I conclude this given 1) the signatures of positive selection shared across such a number of populations and 2) the number of nearly-significant signatures of positive selection in zinc-transporter genes.

Some zinc transporter genes may play a greater role in mediating zinc-associated adaptation amongst global populations. Indeed, a previous study has suggested that zinc-associated adaptation is largely mediated by only a few zinc transporter genes (Roca-Umbert et al. 2022). This study suggested a general enrichment of signatures of positive selection amongst zinc-associated genes (Roca-Umbert et al. 2022), but only explored the signatures of positive selection in metapopulations and individual South Asian populations, using a combination of F_{ST} and haplotype-based methods. However, here, signatures of positive selection are identified at a finer resolution (in individual populations rather than populations grouped as a metapopulation) and use an additional method (*Relate*; more sensitive to the signatures of selection on standing variation than haplotype-based methods, see **Chapter 2**). Hence, this study has greater power to identify more subtle and local adaptation.

I identify the *SLC39A11, SLC30A8, SLC30A10* and *SLC30A1* zinc-transporter genes as frequently sharing signatures of positive selection amongst the same populations, and therefore could represent a common, global network to mediate zinc adaptation. However, strong signatures of positive selection amongst other zinc-transporter genes in individual populations are still observed, as summarised in **Section 4.4.1.1**. Hence, whilst there may be only a few zinc-transporter genes which are largely responsible for zinc-associated adaptation amongst human populations, additional zinc-transporter genes mediate further adaptation, perhaps more uniquely to individual populations.

Indeed, since the zinc-transporter genes carry out a diverse range of biochemical roles within the human body, including structural, regulatory or catalytic roles (Kambe et al. 2015), it is likely that adaptation on some zinc-transporter genes is pleiotropically constrained (Wagner and Zhang 2011; Fraïsse et al. 2019; Mauro and Ghalambor 2020). If pleiotropic constraints vary over zinc-transporter genes (a consideration outside the scope of this study), this could, in theory, result in only a few zinc-transporter genes commonly responding to selective pressures (with other zinc-transporter genes possibly compensating for any resulting biochemical changes). *SLC39A8* has been shown to be highly pleiotropic (associated with Crohn's disease, blood pressure, body mass index and schizophrenia, amongst other traits (Costas 2018)), but the degree of pleiotropy over other zinc-transporter genes remains unclear.

Finally, I turn to exploring the geographic and temporal origin of the inferred shared positive selection on zinc-associated genes. The strong signatures of positive selection shared amongst many non-African populations, partnered with the highly similar non-African haplotypes of *SLC39A11* and *SLC30A9,* suggest that the shared selection event on these genes may have been on an ancestral non-African population, most likely from an allele segregating at low frequency. I propose that this could have been a population migrating out of Africa and living in the Arabian Peninsula (Soares et al. 2012; Haber et al. 2019; Beyer et al. 2021). The Middle-East, especially Iran, is known to have particularly iron and zinc-deficient soils (Ryan et al. 2013), has a history of zinc deficiency disorders (such as severely stunted growth (Halsted et al. 1972; Prasad 2013)) and was the first place where human zinc deficiency was recognised in the 1960s (Halsted et al. 1972; Gibson 2012; Prasad 2013). Adaptation to regulate zinc levels may have thus occurred in an ancestral population living on, and eating from, these deficient soils, and potentially repeated on additional zinc-reporter genes in populations living on elsewhere deficient soils (such as those identified particularly in South Asia, where other populations exhibiting strong signatures of positive selection on zinc-transporter genes reside, *e.g.,* the Makrani (Wessells and Brown 2012; Roca-Umbert et al. 2022)).

The possible exception to this is seen in *SLC39A4*. Previous studies have suggested that proposed adaptation on *SLC39A4*, and its near fixation in West Africa, is due to increased pathogen stress driving lower zinc uptake (Engelken et al. 2014; Zhang et al. 2015a). Indeed, whilst strong signatures of positive selection are identified in *SLC39A4*, these were identified via their degree of differentiation to Yoruba, as calculated by F_{ST}, and only six populations give nearly-significant signatures of positive selection according to *Relate*. There are also fewer uniform haplotypes in non-Africans in comparison to the other candidate zinc-transporter genes. For these reasons, I am more cautious in suggesting that this zinc-transporter was under the same ancestral selective pressures on *SLC39A11* and *SLC30A9*.

4.5.2. Selenium

Somewhat similar to the case of zinc-associated genes, there is a clear network of selenium-associated genes which often share signatures of positive selection over multiple populations: *SGCD, AKAP6, PRKG1* and *KCNMA1.* All four of these genes have intron SNPs associated with selenium regulation (Savas et al. 2010). Moreover, there appears to an epistatic effect between this group of genes, with SNP-SNP interaction indicated between *AKAP6* and *SGCD* and between *AKAP6* and *KCNMA1* (Savas et al. 2010), implying that mutations in these genes may interact to regulate selenium levels, and support their role as an adaptive gene network.

However, other selenium-associated genes show additional signatures of positive selection in individual populations. This, partnered with the frequency of signatures of positive selection on different selenium-associated genes but the relative lack of strong signatures, suggests that adaptation in response to selenium-associated pressures is truly oligogenic in nature (but not extending to polygenic, as investigated in **Chapter 3**). Still, here I focus on the groups of selenium-associated genes co-exhibiting signatures of positive selection amongst the same metapopulations.

PRKG1, AKAP6, SGCD and *EEFSEC* all show signatures of positive selection in East Asian populations, as well as near identical haplotypes shared in individuals of this metapopulation (particularly observed in the haplotypes of *PRKG1*). Many East Asian

populations are known to be living on extremely selenium-deficient soil and hence under selenium-deficiency stress (Xia et al. 2005; White et al. 2015). It is thus possible that this group of genes primarily mediates adaptation in response to selenium-associated pressures in East Asia.

In African populations, who also often live on selenium-deficient soil (Hurst et al. 2013; Ibrahim et al. 2019), the genes exhibiting shared strong signatures of positive selection are *LRP8* and *LHFPL2. LRP8* is a receptor of the selenoprotein P (*SELENOP*), determines the hierarchy of selenium supply to various organs under deficiency (Sarangi et al. 2018) and has been shown to increase mRNA concentrations following *SELENOP* knock-out induced deficiency (Pietschmann et al. 2014). The function of *LHFPL2* is less clear, but a role in selenium metabolism has been suggested by its association with toenail and blood selenium concentrations (identified in a previous genome-wide association study (Cornelis et al. 2015)).

Here, I have identified novel candidate genes for population-specific selenium-associated adaptation. The population differences of the strongest signatures of positive selection are consistent with East Asian and African populations having potentially adapted to selenium stress arising from selenium-deficient soils, but primarily through genetic changes in different groups, or networks, of genes.

4.5.3. Iodine

The iodine-associated genes showing strong signatures of positive selection are the least shared amongst populations in comparison to the other micronutrients examined here, and I suggest that potential adaptation to iodine is more focused to individual populations, rather than shared amongst populations. However, here I outline one key exception.

The Maya population of the Americas and the Mbuti population of Central Africa both share signatures of positive selection in four iodine-associated genes, three of which are thyroid-receptors (*THRA, THRB, TRIP4*). Given the high levels of goitre in modern Mexicans (Kelly and Snedden 1960), it is possible that the Maya experienced a low-iodine environment in their ancestral home. Rainforest populations, including those of Central Africa, are also known to be living on soils deficient in iodine (Cifor 2006) and the Mbuti may have been exposed to a similar iodine deficiency selective pressure. What is most intriguing here, is both populations' distinctive short stature (height < 160cm (Perry and Dominy 2009)) which may be mediated by these three thyroid-receptors (Rose 1995; Moran and Chatterjee 2015; Xu et al. 2016). A substantially lower rate of goitre is recorded in the short-statured Efe population compared to neighbouring Bantu-speaking populations (Dormitzer et al. 1989), where both populations live in similarly low-iodine soils, presenting a further link between short stature and resistance to iodine deficiency. Hence, the characteristically short stature of these populations may be tightly linked to iodine metabolism.

Other populations exhibiting signatures of genetic adaptation in iodine-associated genes include the Palestinian (signatures of positive selection bypassing the most stringent threshold in the *THRB* gene) and the Uygur population (where multiple iodine-associated genes exhibit signatures of positive selection). This latter population live in the now Xinjiang Uygur Autonomous Region of Northern China, which includes areas of severely iodine deficient soils (likely as a result from to its distance from the ocean (Yang et al. 2021)). Urine iodine levels have been shown to be significantly different in Uygur and

Han Chinese pregnant women (Renaguli et al. 2018), suggesting a potentially different metabolic reaction to iodine in these populations.

4.5.4. Calcium

Multiple populations also exhibit signatures of calcium-associated adaptation, inferred by multiple genes showing strong signatures of positive selection. These include the Biaka and Bantu-speaking populations of Africa, the She-Miao-Tujia and Japanese of East Asia, the Kalash of Central-South Asia and the French of Europe. However, it is unclear if the putative calcium-associated adaptation could be due to cultural or environmental factors, (*i.e.,* cultural differences in diet or underlying soil composition), since data on the calcium levels in these soils is sparse.

The strongest signature of positive selection of this entire thesis (see **Chapter 3**) is observed in the calcium-associated *ATP2B2* gene of the African Mandenka. It has previously been shown that Gambian populations, many of whom have Mandenka ancestry, have low calcium urinary excretion under low calcium intake, which has been suggested to maintain bone health even under low-calcium diets (Aspray et al. 2005; Redmond et al. 2015). The Mandenka live throughout West Africa (countries with the highest number of individuals with Mandenka ancestry include Senegal, Mali, Guinea and The Gambia (Currat et al. 2002)) and whilst there has been some indication of reduced calcium in some West African soils, this has not yet been linked explicitly to the Mandenka population (Issaka et al. 1996; Baumann et al. 2021).

However, strong signatures of positive selection have been identified on this gene amongst the majority of populations in this study. *ATP2B2* is thus proposed as strong candidate for having undergone widespread adaptation in modern humans (rather than isolated to the Mandenka population), most likely as a result of selection on standing variation. Still, as outlined in **Chapter 3**, it is difficult to confidently associate these signatures with calcium-associated selective pressures, given the role this gene plays in many human diseases.

I investigate the timing of proposed selection on this gene and four other calcium-associated genes (*ATP2B4*, *SLC8A1, SLC8A2* and *SLC8A3)* by combining the inferred log-likelihood ratios of selection and the allele trajectories of focal SNPs. For the majority of populations with signatures of positive selection, I suggest that the onset of selection was approximately 40kya and hence could reflect selection in an early non-African population. Given the lack of soil data for calcium levels, including that of soils in the Arabian Peninsula, it is not possible to confidently link these genomic signatures to selection in a migrating Out of Africa population, although it is worth noting that the majority of the results do not support selection accompanying Neolithic changes to the diet.

The exception to this is observed in the Mozabite population; the high log-likelihood ratio of selection and frequency trajectories of the focal SNPs of *ATP2B4* and *SLC8A3* suggest that selection could have started approximately 10 – 5kya. Both of these genes control calcium transport, with the former associated with the calcium absorption of laying hens (Gloux et al. 2019), and it is thus possible that adaptation in response to calcium in the diet occurred in the Mozabite population. This may be a result of Neolithic dietary changes surrounding this time point. This agrees with previous studies suggesting calcium adaptation in the Mozabite (Hughes et al. 2008). Finally, very recent increases in frequency of the focal SNP of *SLC8A3* are inferred in the American Pima population (between 10 – 5kya). Interestingly, and in agreement with calcium-associated selective

pressures in this population, the traditional diet of this population is high in iron but low in calcium (Greenhouse 1981).

4.5.5. Iron

Finally, iron-associated genes also show signatures of positive selection that appear to co-occur with the timing of major human migrations, albeit possibly a little more recent than for calcium. For the iron-associated genes of interest, steep increases in allele frequency occurring at 30 – 20kya are inferred, broadly agreeing with the estimated log likelihood ratios of selection estimated over time. The more recent inferred onsets of selection compared to calcium-associated genes, alongside the geographical partitioning of strong signatures of positive selection (such as those unique to East Asian populations in the *RHOA* gene), also suggests that such adaptation may have been driven by the novel environmental conditions of different Eurasian environments, rather than a common environment encountered immediately following the exit from Africa. By extension, adaptation in response to iron levels is also more likely to be largely driven by novel environmental pressures, such as soil levels, rather than recent changes in diet.

The signatures of positive selection identified in the iron storage protein *FTMT*, which plays a central role in protecting mitochondria from iron excess (Levi et al. 2021), suggest selection around ~30 – 20kya in the Yakut population (inhabiting modern day Siberia). Increases of frequency of the focal SNP of the *ARHGEF3* gene, which regulates iron intake and erythroid cell maturation (Serbanovic-Canic et al. 2011), are also inferred in the American Maya population at this time (estimated as ~20kya). This latter time is the approximate time of the stasis of ancestral American populations in the Bering Strait (Raghavan, DeGiorgio, et al. 2014; Raghavan, Skoglund, et al. 2014). Hence, iron-associated adaptation may be driven by environmental factors experienced by populations living in the Siberia. In the absence of data on the iron content of these soils this can only be speculated.

Previous literature has suggested iron-associated adaptation in European populations (Distante et al. 2004; Ye et al. 2015), of which some support is presented here. The *HIF1A* gene especially, a hypoxia-inducible factor that plays a role in iron homeostasis (Shah and Xie 2014) exhibits both strong signatures of positive selection and very recent inferred frequency increases (between 10 – 5kya) in the Basque population of Europe. Notably, this is a novel target of iron-associated selection considering the existing literature. Finally, in support of the micronutrient-associated adaptation in the Mozabite surrounding the time of the Neolithic transition and associated dietary changes, there are signatures of positive selection and a striking increase of allele frequency of the focal SNP of the *FTMT* gene between 10 – 5kya in the Mozabite population. However, the signatures of positive selection identified in Mozabite this study may be conflated with those arising from admixture of this North African population with Europeans (Hughes et al. 2008), and this would benefit from further investigation.

4.5.6. Strengths and Limitations

As informed from previous work (**Chapter 3**), I identify the micronutrients with the strongest evidence of acting as a selective driver in human evolutionary history and investigate the geographic distribution and strength of their associated signatures of positive selection in populations across the globe. I consider co-occurrence of signatures, haplotype diversity and inferred allele frequency over time to better elucidate the

evolutionary history of genes within five micronutrient-associated gene sets, and suggest the most likely origins of putative selection, in both space and time. By exploring the signatures of positive selection on only five micronutrient-associated gene sets, I am able to focus the analysis to a degree not possible in **Chapter 3**, identifying similarities and differences between the genomic adaptation of different human populations, and use computationally intensive methods (*e.g., CLUES* (Stern et al. 2019)) to provide further support of and information on putative adaptation scenarios.

Whilst the focus on only five micronutrients in this chapter removes some limitations of the comprehensive approach of **Chapter 3**, some remain. As addressed in **Chapter 3**, there is limited information on soil information or contemporary health data which, if available, would more clearly allow an evaluation on the link between the micronutrient content in the diets of ancestral populations and putative signatures of micronutrient-associated adaptation. Also, each SNP identified as having a signature of positive selection is not necessarily a true target of selection, and the methods to identify selection will naturally result in both false positives and false negatives. To mediate this, I consider of signatures of positive selection at the gene set level and evaluate the presence of nearly-significant signatures of positive selection.

Finally, the exact functional role of micronutrient-associated genes must be considered. Many of the micronutrient-associated genes explored here are responsible for many functions in the human body, and signatures of positive selection identified may not be strictly related to micronutrient metabolism. There may also be multiple stressors which have resulted in the identified adaptive signatures, as suggested in the case of zinc-associated adaptation (*e.g.,* driven by soil levels and/or pathogen resistance). I can thus only suggest the most likely micronutrient-associated drivers according to inferences of time of selection (*e.g.,* if they most closely coincide with migrations to novel environments, and therefore potential soil-related stress, or the Neolithic transition, and therefore cultural changes to diet) but confidently teasing apart the individual selective drivers of these cases of proposed adaptation will take significant further work and functional analysis, and may not be easily summarised across large numbers of populations.

4.5.7. Summary

When signatures of positive selection coincide with soil and public health data, as is the case for some of the strongest examples here, there is good reason to suggest that they likely underlie adaptation in response to micronutrient levels. Whilst contemporary public health data for micronutrient deficiencies does not necessarily reflect underlying soil levels and is often tightly linked to national food or economic inequality (Shenkin 2006; Bhutta and Salam 2012; Bailey et al. 2015), and modern-day soil levels are not necessarily indicative of the ancestral environment (owing to farming and other agricultural practices (Shahid et al. 2018; Dhaliwal et al. 2019; Alewell et al. 2020)), I present evidence that soil composition has contributed to driving population-specific adaptation to micronutrient levels, alongside potential additional cultural dietary drivers.

I particularly propose that Middle-Eastern geology has driven micronutrient-associated adaptation in an ancestral non-African population, hence playing an important role in shaping the genomic diversity of many modern human populations. The role of environment-induced adaptation in Middle-Eastern populations is considerably understudied considering the likely importance of this environment in driving adaptations in populations migrating Out-of-Africa, and consequent likely implications

(including those possibly surrounding differential health outcomes, see **Chapter 1**) for all non-African populations. Here, I outline one example of how this environment may have shaped non-African genomic diversity, but suggest that many more exist and warrant significant further study.

I also propose cases where populations have responded to the same micronutrient-associated stresses in soil levels via adaptation of the same or similar genes (as suggested in the adaptive response to high elevation in human populations (Foll et al. 2014; Huerta-Sánchez et al. 2014; Ilardo and Nielsen 2018)). Most notably, I propose this in the response of Maya and Mbuti populations to iodine-deficient soils, and further suggest this as a causal link to short stature in these populations. I also identify cases where populations have responded to the same micronutrient-associated stresses via adaptation of different sets of genes, most notably the response of many African and East Asian populations to selenium-deficient soils. Finally, there is limited evidence that the dietary changes in the Neolithic drove widespread calcium or iron-associated adaptation, but there are some cases where selection on calcium or ion-associated genes is inferred to be more recent, coinciding with major dietary changes in human history.

To more comprehensively understand micronutrient-associated adaptation, future studies would benefit from a deeper understanding of ancestral soil environments and differences in the prevalence of micronutrient-associated pathologies of modern populations according to ancestry. Alongside functional analysis on candidate genes regarding their role in their associated micronutrient regulation or metabolism, this will more clearly pinpoint the selective drivers of such adaptive signatures.

4.6. Conclusion

Here, I build on previous work (**Chapter 3**) to more thoroughly investigate the signatures of positive selection identified on the genes associated with five micronutrients: zinc, calcium, selenium, iron and iodine. I identify the groups of micronutrient-associated genes which have the strongest evidence for mediating micronutrient-associated selective pressures in human populations across the globe, and suggest that populations may evolve through different genomic routes to the same micronutrient-associated pressures. I give evidence for older selection events in ancestral non-African populations, particularly in zinc-associated adaptation in the Middle-East, and present a small number of micronutrient-associated adaptation events that more likely surround the Neolithic dietary transition.

Chapter 5: Ancient Loss of Catalytic Selenocysteine Spurred Convergent Evolution in a Mammalian Oxidoreductase

This chapter is based upon the work undertaken in the preprint: ***Ancient loss of catalytic selenocysteine spurred convergent adaptation in a mammalian oxidoreductase*** (Rees et al. 2023)*. Where specified, some work was primarily undertaken by collaborators.*

5.1. Overview

Catalytic residues are often conserved in proteins, with mutations that occur at or close to key sites frequently reducing catalytic activity and corresponding fitness of the enzyme (Sharir-Ivry and Xia 2021). When such deleterious mutations persist, they often demonstrate evolutionary trajectories which either recover catalytic function or open new protein functions (Jensen 1976; Gromer et al. 2003; Jayaraman et al. 2022). Here, we investigate the evolutionary and functional trajectories that follow the loss of the key catalytic residue in a mammalian oxidoreductase.

Selenocysteine (Sec), the 21st amino acid specified by the genetic code, is a rare selenium-containing residue found in the catalytic site of selenoprotein oxidoreductases. These proteins mediate the essential biological effects of the rare trace element selenium (explored in terms of its role in modern human health and adaptation in **Chapters 1, 3, 4**). Sec is analogous to the common cysteine (Cys) amino acid but its selenium atom offers physical-chemical properties not provided by the corresponding sulfur atom in Cys. Hence, exchanges of Sec to Cys in the catalytic sites of vertebrate selenoproteins are often under strong purifying selection (Castellano et al. 2009). Whilst the presence of both Sec and Cys orthologues are rare, these are observed in Glutathione Peroxidase 6 (GPX6), which has independently exchanged Sec for Cys less than one hundred million years ago in several mammalian lineages.

We reconstructed and assayed ancient GPX6 enzymes before and after the loss of Sec, alongside the modern mouse protein, and found them to have lost their classic ability to reduce hydroperoxides using glutathione (GSH). This loss of function, however, was accompanied by additional amino acid changes in the catalytic domain, with protein sites showing signatures of adaptive convergence across distant lineages abandoning Sec in GPX6. This demonstrates a narrow evolutionary path when sulfur in Cys impairs catalysis, with pleiotropy and epistasis likely driving the observed convergent evolution and triggering enzymatic properties beyond those in classic GPXs.

5.2. Background

Selection does not act equally over all sites of a protein; those sites with functional importance, or contributing to the stability of a protein, are under stronger selective pressure to conserve their vital roles (Sharir-Ivry and Xia 2021). Catalytic residues, those that lower the activation energy of reactions and thereby increase enzymatic turnover, are a key example of such largely conserved sites, and show slower rates of evolution compared to other sites within a protein. Catalytic sites have even been show to exert a

strong gradient of conservation on nearby sites, such that the degree of conservation on a site increases with the closeness to the key catalytic residue (Sharir-Ivry and Xia 2021).

Mutations that occur in these evolutionary constrained active sites, or indeed in nearby sites, typically reduce catalytic activity (a proxy for fitness in enzymes; (Carter and Wells 1988; Loeb et al. 1989; Rennell et al. 1991)) and are frequently removed by purifying selection. Still, these deleterious mutations may occur and persist, with their effects often mediated by compensatory mutations that restore catalytic ability, and therefore fitness of an enzyme (Jensen 1976; Gromer et al. 2003; Cha et al. 2013). Such compensatory mutations, often occurring near the deleterious mutation, improve the fitness of a protein when accompanying a deleterious mutation but are otherwise neutral or even slightly deleterious (Davis et al. 2009).

Therefore, such compensatory changes reflect a specific form of intragenic epistasis, whereby they increase the fitness of a deleterious mutation to either become neutral or advantageous, and increase the possibility of its fixation in the population (Davis et al. 2009; Jayaraman et al. 2022). These compensatory mutations may either precede or follow the deleterious mutation event, restoring fitness or effectively preventing the loss of fitness on the onset of the deleterious mutation (Jayaraman et al. 2022), and their onsets may be considerably spread around the appearance of the deleterious mutation (Jayaraman et al. 2022). Whilst the evolutionary landscape of such mutations is therefore highly complex, the role of compensatory mutations in recovering fitness has been implicated across a wide range of biological scenarios, including following fixation of deleterious mutations in small populations, restoring antibiotic or pesticide resistance, and repairing ancestral catalytic ability (Jensen 1976; Gromer et al. 2003; Whitlock et al. 2003; Maisnier-Patin and Andersson 2004; Cha et al. 2013; Larsson and Flach 2022).

In other cases, deleterious mutations may prompt evolutionary trajectories that open the protein to novel functions (Jensen 1976; Gromer et al. 2003; Covert et al. 2013). Whilst gene duplication often precedes the appearance of new adaptations, which often evolve as a result of one or both gene copies being released from their previous functional constraint (Hughes 1997), deleterious mutations may also allow adaptations that were previously unavailable by way of interacting with a conditionally beneficial mutation (Lenski et al. 2003; Covert et al. 2013). Indeed, this has been implicated in the evolution of many novel enzymes and their functions, including cystallins of the eye, isocitrate dehydrogenase of the Krebs cycle and novel organophosphorus hydrolase activity mediating insecticide resistance (Piatigorsky and Wistow 1991; Dean and Golding 1997; Newcomb et al. 1997).

As outlined, protein evolution following a deleterious mutation at their active site depends heavily on intragenic epistasis, whether that is an evolutionary trajectory which repairs original function or helps traverse fitness space to develop a novel function. Mutational trajectories are limited by the enzyme's sequence (with pleiotropy further limiting trajectories that improve one enzymatic property but compromise another (Weinreich et al. 2006; Storz 2016)). This is best represented in orthologous proteins, whose sequence conservation among species provides similar genetic backgrounds to mutations (Lunzer et al. 2010; Shah et al. 2015). Ultimately, this can result in narrow fitness trajectories of such similar proteins, and give rise to convergent, or parallel, changes across closely related lineages (Weinreich et al. 2006; Storz 2016).

5.2.1. Selenoprotein Evolution

Here, we investigate the loss of a key catalytic residue in the selenoprotein Glutathione Peroxidase 6 (GPX6). In this protein, there has been sporadic replacement throughout mammalian history of selenocysteine (Sec) with cysteine (Cys) at the catalytic site. This is expected to result in an immediate loss of catalytic ability, as previously shown for this amino acid exchange (Axley et al. 1991; Berry et al. 1992; Lee et al. 2000; Johansson et al. 2005; Arnér 2010; Kim et al. 2015; Reich and Hondal 2016). Since this protein is unique in its family for containing both Sec and Cys as its key catalytic residue in contemporary mammals, with other GPX proteins exclusively containing either Sec (GPX1, 2, 3 and 4) or Cys (GPX5, 7 and 8; (Mariotti et al. 2012), it presents good opportunity to infer the evolutionary trajectories that follow a deleterious mutation at a unusual catalytic site, in direct comparison to the orthologues without such a mutation.

Sec is the 21st amino acid and the defining catalytic residue of selenoproteins, a family of proteins that uses and mediates the biological effects of the rare trace element selenium. Selenium is an essential micronutrient in many organisms and is responsible for a wealth of vital biochemical functions (Labunskyy et al. 2014). It is particularly associated with development, immune response and reproduction (Köhrle 2000; Rayman 2012), and deficiencies in humans result in a range of pathologies, including those outlined above and, in extreme cases, heart and bone diseases such as those endemic to selenium-deficient areas of China (Xia et al. 2005).

The selenium-containing amino acid Sec is unusually encoded by a UGA stop codon, and its insertion requires a Sec insertion sequence (SECIS) element to redefine this codon to specify Sec insertion (Berry et al. 1992). This stem loop structure is in the 3'UTR of the mRNA in selenoproteins in mammals, as well as all other eukaryotes and archaea (Labunskyy et al. 2014). Selenoproteins using such a molecular structure are rare, with only 25 selenoproteins making up the selenoproteome in humans (Kryukov et al. 2003). This number is mostly conserved in mammals (Mariotti et al. 2012), but shows a general decrease in non-mammal organisms (*e.g.,* only 3 selenoproteins in *Drosphila melanogaster* (Castellano et al. 2001)), with the exception of aquatic organisms which often have a larger selenoproteome (Lobanov et al. 2007).

Sec is often a key catalytic residue at the active site of enzymes and plays a key role in catalytic redox reactions, including reductions of thioredoxin, activation and inactivation of thyroid hormones, repairing oxidised methionines in proteins and removal of hydroperoxides (the latter as in the GPX family (Santesmasses et al. 2020)). Whilst Sec's role is often considered unique, many selenoproteins have been found with this catalytic residue entirely replaced by Cys (UGU or UGC), the analogous amino acid containing a sulfur-containing thiol group in place of the selenium-containing selenol group of Sec (Stadtman 1996).

However, such Sec-to-Cys substitutions across orthologous selenoproteins, as seen in mammalian GPX6, are rare (Castellano et al. 2005). There is a low exchangeability of Sec and Cys in catalysis, where Cys displays lower catalytic activity, nucleophilicity and efficiency as a leaving group when compared to Sec (Axley et al. 1991; Berry et al. 1992; Lee et al. 2000; Johansson et al. 2005; Arnér 2010; Kim et al. 2015; Reich and Hondal 2016). Hence, the exchange of Sec to Cys is often deleterious and strong purifying selection limits these exchanges in nature. When exchanges between Sec and Cys do occur and become fixed, these are often following gene duplications that may release the

duplicated gene from its catalytic functional restraint (Mariotti et al. 2012; Magadum et al. 2013).

Indeed, all Cys-containing proteins of the vertebrate GPX family are a result of duplications in early history: $GPX5_{Cys}$ from $GPX3_{Sec}$ duplication (around 300 Mya); $GPX8_{Cys}$ from $GPX7_{Cys}$ or $GPX4_{Sec}$ duplication (likely 450 Mya); $GPX7_{Cys}$ from $GPX4_{Sec}$ duplication (more than 1,000 Mya) (Hedges 2002; Castellano et al. 2009; Trenz et al. 2021).

5.2.2. Study Overview

The presence of both the Sec and Cys-containing orthologues of GPX6 is therefore highly unusual, particularly in vertebrates, and allows us to ask the immediate evolutionary response to such an exchange. We first consider if, in view of the deleterious nature of losing Sec, the exchange between Sec and Cys results in the emergence of compensatory mutations that act to repair catalytic ability, as demonstrated by (Gromer et al. 2003) in *Drosophila*, and if these compensatory mutations are shared over all Cys-containing mammalian lineages.

We also consider if the functional pathway of $GPX6_{Cys}$ changes as a result of the exchange of its key catalytic residue. Whilst GPX proteins, which contain either Sec or Cys at their defining catalytic site, all protect the cell from oxidative damage (Tosatto et al. 2008), they do so via different pathways. Classic GPX_{Sec} activity reduces hydroperoxides, particularly hydrogen and lipid peroxides, with glutathione (GSH) as a cofactor (Trenz et al. 2021). GPX_{Cys} proteins, on the other hand, have evolved a preference for other cofactors, for example thioredoxin in $GPX5_{Cys}$ or protein disulfide isomerase (PDI) in $GPX7_{Cys}$ and $GPX8_{Cys}$ (Nguyen et al. 2011). These Cys-containing proteins not only act on alternative substrates for peroxidation but may also have additional functions, including signalling and oxidative protein folding (Nguyen et al. 2011; Taylor et al. 2013; Buday and Conrad 2021). We hence also ask if the Cys-containing orthologues of GPX6 also develop novel functional pathways, on account of their lower catalytic turnover.

By reconstructing GPX6 protein evolution throughout mammalian history, we are first able to identify five independent losses of Sec in mammals, surrounded by a burst of amino acid changes in the catalytic domain. An unusual number of the amino acid changes that accompany Sec loss are shared across distant lineages, indicating a narrow evolutionary path, likely mediated by pleiotropy and epistasis, available to proteins when the sulfur-containing Cys impairs catalysis. We also reconstruct and assay ancient enzymes before and after Sec loss in the *Eumuroida* lineage, and find them to have lost their classic ability to reduce hydroperoxides using glutathione (GSH). Hence, such a narrow evolutionary path seems to trigger enzymatic properties beyond those in classic GPXs, reappraising function rather than recovering previous catalytic ability. Thus, these findings are an unusual example of adaptive convergence towards unexplored oxidoreductase functions during mammalian evolution.

5.3. Methods

5.3.1. GPX Sequences

The GPX6 coding sequences and proteins for 22 present-day mammal species were obtained from *SelenoDB* 2.0 (now available at selenodb.crg.eu; (Romagné et al. 2014)) and *Ensembl* (Yates et al. 2020), chosen for their availability and breadth across the

mammalian tree (**Table S5.1**).The *Ensembl* species tree (available at www.ensembl.org) was used to give the phylogeny of these mammals with the exception of the walrus, which was added according to various additional sources (Higdon et al. 2007). These species include nine mammals where GPX6 contains Cys in the place of Sec.

The orthologous GPX6 coding sequences and proteins were aligned using *MAFFT* (Katoh et al. 2019). The posterior probability of each individual aligned position was then calculated using a modified version of *HMMER* (Potter et al. 2018), which first converts each protein multiple alignment into a Hidden Markov Model before using a forward-backward algorithm to perform posterior decoding (Durbin et al. 1998). The calculated posterior probability integrates the uncertainty of the alignment around an aligned position, representing our degree of confidence in each individual aligned protein residue or gap in a multiple alignment.

Positions with an average posterior probability below 0.95 were then removed, due to concerns of misalignment, and not included in further analysis using *PAML* (Yang 2007). The removed positions are, in general, found surrounding gaps or points of sequence divergence, which both contribute to alignment uncertainty. Nevertheless, our probabilistic approach allowed us to keep regions containing gaps or amino acid differences that were confidently aligned in the multiple alignments.

The coding sequences for other members of the GPX family (four GPX proteins where all species have Sec (GPX1, 2, 3 and 4) and three GPX proteins where all species contain Cys (GPX5, 7 and 8)), were also obtained from SelenoDB 2.0 (Romagné et al. 2014) or, if not available, from *Ensembl* (Yates et al. 2020) (**Table S5.1**). These proteins are aligned following the methodology as described above.

5.3.2. Ancestral Reconstruction of GPX Proteins

5.3.2.1. Inferring the Loss of Sec

The ancestral sequences of GPX6 for our set of 22 mammals were reconstructed using the mammalian tree of these species, their present-day sequences and the *PAML* package (Yang 2007). We used this package to infer the sequence of all ancestral nodes across the mammalian tree and pinpoint the inferred independent losses of Sec throughout the mammalian lineage. The independent losses of Sec within the lineages leading to the walrus and cat were inferred according to the most parsimonious scenario when accounting for the presence of Sec in the Ursidae lineage (included as bear in **Fig. 5.1**). The approximate ages of lineages with Sec loss are collected from various sources describing split times in the mammalian phylogeny (Huchon et al. 2002; Steppan et al. 2004; Higdon et al. 2007; Hallström and Janke 2008; Chatterjee et al. 2009; Nyakatura and Bininda-Emonds 2012).

Further to the *PAML* inference, we also inferred the ancestral sequences using two additional programs: *Ancestor* v1.1 (Diallo et al. 2010) and *FastML* (Moshe and Pupko 2019). *FastML* has options to use either amino acid or nucleotide sequences of contemporary species as input to infer the ancestral sequences, whereas *Ancestor* v1.1 (alongside *PAML*) only uses the nucleotide sequences for inferences. Hence, using both *FastML* input methods, this gives four inferred sets of sequences for all ancestral nodes. The four inferred sequences were then aligned using *MAFFT* (Katoh et al. 2019) and the residue with the most support was taken as the consensus residue for each site.

5.3.2.2. Ancestral Proteins Along the *Eumuroida* Lineage

Following our inference of ancestral sequences, we were then able to reconstruct three ancient proteins along the *Eumuroida* lineage, where *Eumuroida* includes rats, mice and closely related rodents (see **Figure 5.1**). These proteins are: 1) the protein just prior to the loss of Sec in the ancestor of *Eumuroida* (Eu-GPX6$_{Sec}$); 2) the same ancestral protein but with Sec exchanged for Cys (Eu-GPX6$_{Cys}$); and 3) the protein at the derived end of the *Eumuroida* branch, now containing the additional 25 sites that have changed along the *Eumuroida* branch (Eu-GPX6$_{Cys+25}$).

As previously described, the residue with most support from the four inferred sequences was taken as the consensus residue for each site, with the exception of site 54 in Eu-GPX6$_{Cys+25}$. Here, the consensus residue was taken as "Q" (Glutamine, CAA) despite the methods used suggesting "H" (Histidine, CAT or CAC) since "H" is not present at this site for any of the contemporary species. Of the 217 amino acid sites, 208 (95.85%) were resolved unanimously across the four inference methods. Of the remaining nine sites that were inferred differently across the methods, seven (3.23% of total sites) of these sites differed across the inference of the Eu-GPX6$_{Sec}$ protein and two (0.92% of total sites) differed across the inference of the Eu-GPX6$_{Cys+25}$. We use these consensus sequences to provide the final Eu-GPX6$_{Sec}$ and Eu-GPX6$_{Cys+25}$ proteins, with sites calculated as having an average posterior probability below 0.9 (as calculated using *HMMER* (Potter et al. 2018)) removed from subsequent *PAML* analysis.

5.3.3. Inferring Rate of Evolution

We use the dN/dS ratio as a quantification of the rate of evolution and strength of selection acting on proteins, where dN is the rate of non-synonymous substitutions per non-synonymous sites and dS is the rate of synonymous substitutions per synonymous sites. All dN/dS ratios were computed using the *CODEML* package from *PAML* (Yang 2007), using the aligned GPX6 coding sequences and mammalian tree topology. The UGA codon encoding the Sec amino acid was considered an ambiguity character and not included in the dN/dS calculation, hence making our calculations conservative when comparing the rate of evolution in proteins that have exchanged Sec for Cys to those who have maintained Sec.

5.3.3.1. dN/dS Ratios in GPX Proteins

We first calculated independent dN/dS ratios for each branch in the GPX6 mammalian phylogeny using the free-ratio model (model=1) in *PAML*. This allows the dN/dS ratio to vary amongst the branches of the phylogenetic tree and was used to compare the rate of evolution in the lineages that retain Sec and those that have exchanged Sec for Cys. Given this preliminary comparison, the *CODEML* branch model (model = 2) was then used to explicitly test our hypothesis of a faster rate of evolution in lineages where Sec was lost.

The *CODEML* branch model (model=2) allows us to specify the number of independent dN/dS ratios across set groups of branches. We used this model to compare the dN/dS ratios between three groups of branches: the branches with Sec (**Fig 5.1**; solid red branches), the branches where Sec is exchanged for Cys (**Fig 5.1**; dashed green branches) and the branches where Cys is maintained (**Fig 5.1**; solid green branches). Hence, we ask if the dN/dS ratio was significantly different in lineages at the time surrounding the loss

of Sec compared to lineages where Sec was not lost, or where Cys was maintained (under the assumption that any fitness reduced as a result of the loss of Sec had since been recovered).

Once dN/dS ratios across the three groups of branches had been calculated, we compared this branch model to the null model (M0 model, model=0), which estimates a singular dN/dS value for all branches. We compare the likelihood of each of the two models to give a likelihood ratio, which was used to calculate the significance of the difference in fit between the two models in the form of a $pvalue$. Hence, we explicitly ask if three dN/dS ratios across the tree is a significantly more likely fit than the null model of a singular dN/dS ratio across all branches.

We repeated this analysis for all other genes in the GPX family, comparing dN/dS ratios calculated over the three groups of their analogous branches (those analogous branches that have lost Sec, maintained Sec or maintained Cys in the GPX6 phylogeny) to the null model of one dN/dS over the entire phylogeny.

5.3.3.2. dN/dS Ratios in Protein Domains in GPX Proteins

We then separated the protein into its three domains: N-terminus, GPX domain and C-terminus (as defined in the $PFAM$ database (Mistry et al. 2021)) to further explore how evolutionary rates may vary over the protein. Of these three domains, the GPX domain is considered essential for the catalytic activity of the enzyme, alongside the C-terminus which also contributes to catalytic function (Toppo et al. 2008). We repeated the analysis outlined above separately for each of the three domains of GPX6, as well as for the three domains of each of the additional GPX genes.

5.3.3.3. dN/dS Ratios in GPX3

We found an additional two GPX proteins unexpectedly lacking the Sec residue: GPX3 in both the Hoffman's two-toed sloth and the kangaroo rate. Here, the Sec has been exchanged for either glutamine (in the case of the sloth) or for serine (in the case of the kangaroo rat). Because of these exchanges, we removed the sloth-GPX3 and kangaroo rat-GPX3 from the branch model analysis of dN/dS rates in GPX3 (but maintained in the following branch-site analysis, see **Section 5.3.4**).

5.3.4. Inferring Selection on the GPX6 Sites

We used the Site model in $PAML$ (Yang 2007) to test for selection acting on individual sites across the entire tree, comparing model 7 (beta; model = 0 , NSsites=7) to model 8 (beta plus selection; model =0, NSsites=8). Here, model 7 is the null model of a beta distributed variable selective pressure across sites, whereas model 8 is the beta distributed model plus positive selection. Given that there was a significant difference between these models, we then tested for selection acting on sites in the GPX domain along specified branches across the tree. We used the Branch-Site model (model=2, NSsites=2) to calculate the probability of each site of foreground branches (as specified in the model) being under selection according to $PAML$'s Bayes Empirical Bayes (BEB) inference method (Yang et al. 2005).

This model allows the dN/dS ratio to vary both amongst sites and amongst the specified foreground and background branches, classifying the sites into those that have dN/dS

values that remain the same on the foreground and background branches ($\omega < 1$ or $\omega = 1$ in both branches) and those that differ amongst the branches ($\omega < 1$ or $\omega = 1$ in background branches and $\omega > 1$ in foreground branches), outputting the proportion of each site class. This method then calculates the posterior probability of each site being under selection in the foreground branches, whilst accounting for sampling errors by using a Bayesian prior (Yang et al., 2005). This model is compared to the corresponding null model, which is the same in all ways apart from the fixation of ω_2.

We first use this model to infer the probability of selection acting on sites along the branches where Sec was inferred to be lost for the GPX domain only. Having found significant evidence for selection on particular sites within this region, we then extended this model to test along the entire protein region for the same foreground branches. Given this test yielded a non-significant result, we repeated the model to test for selection shared on sites along the entire protein region on the most closely related branches.

Since significant evidence is observed for selection acting on the same sites across the more closely related lineages where Sec was inferred to be lost (the branch leading to squirrel monkey-marmoset (Cys-primate branch), the *Eumuroida* branch and the branch leading to rabbit; see **Fig. 5.1**), we then test if these probabilities are enriched in certain subsets of sites using Mann-Whitney U tests.

5.3.5. Identifying Convergent Changes

5.3.5.1. Convergence Across Cys-branches

Convergent changes in GPX6 across lineages were identified using $CONVERG2$ (Zhang and Kumar 1997). The definition of convergent amino acid changes used here includes sites that have changed from a different ancestral amino acid to the same derived amino acid and sites that have changed from the same ancestral amino acid to the same derived amino acid (other studies may refer to these as parallel changes, see **Section 1.2.2**).

Convergent changes were identified between the $GPX6_{Cys}$ lineages: either the branches where Sec was exchanged for Cys or the species branches where Cys was maintained. The observed frequency of these convergent changes was then compared with the expected frequency of convergent changes, also calculated using $CONVERG2$.

Since the pathway to recover catalytic activity may not be limited to the same amino acid changes but still may be restricted to particular sites in the protein, we edited the $CONVERG2$ programme to also identify convergent site changes which do not result in the same amino acid across branches (hence, simply identifying sites that show repeated amino acid changes across lineages). Such identified sites are also included in our definition of convergent sites hereinafter.

Where the sequences for the species containing Cys in GPX6 were available, the equivalent analyses were run on the Sec-containing GPX proteins (GPX1, 2, 3 and 4) and the Cys-containing proteins (GPX7 and 8). We advise that the focus should be on the convergence results for GPX3 and GPX5 for two reasons: 1) these proteins are the immediate paralogues to GPX6; and 2) the gaps in the other proteins do not allow, we believe, a full representation of the potential instances of convergence.

5.3.5.2. Simulating Expected Convergence

The evolution of the GPX6 protein sequence across our mammalian phylogeny was simulated using $Seq - Gen$ (Rambaut and Grassly 1997). This simulation begins with the

inferred ancestral sequence at the base of our mammalian clade and runs until all modern mammalian proteins are evolved, using the JTT model of amino acid substitution (Jones et al, 2008). Tree lengths were given by the rate of amino acid changes along each branch of the mammalian tree as from the calculated dN value in the $CODEML$ package from $PAML$ (Yang 2007). Hence, the simulation recreates chance amino acid exchanges along each branch at its observed rate.

Each simulation was run 1,000 times and, for each simulation run, $CONVERG2$ (Zhang and Kumar 1997) was used to identify convergent site changes between the lineages where Sec was lost for Cys. The distribution of convergent changes under this expected rate of amino acid exchange is then plotted, and compared to the observed number of convergent site changes. Equivalent simulations were run for all other GPX proteins, and we further compared the observed and expected number of convergent site changes for these proteins.

To confirm that the higher number of observed convergent changes relative to our expectation are focused within the functional GPX domain, and that are conclusions aren't simply an artefact of elevated evolutionary rate, we also repeated these simulations on only this domain (tree lengths given by the rate of amino acid changes from the GPX domain only).

5.3.5.3. Selection Across Convergent Sites

We asked if the convergent sites are enriched for posterior probabilities of selection by comparing the posterior probabilities of selection acting on convergent sites to the remaining sites in the GPX protein. Here, we use the posterior probabilities of selection as calculated acting on sites in the branches leading to Cys-primate branch, the *Eumuroida* branch and the branch leading to rabbit (BEB results of the Branch-Site Model, see **Section 5.3.4**) and exclude convergent sites only identified using either the cat or walrus terminal branches, since they are excluded from the probability calculation. To test for enrichment, we use a Mann-Whitney U test to account for the non-parametric data.

5.3.5.4. Convergence Across *Eumuroida*

Given that the highest level of convergence is identified between the basal *Eumuroida* and it's genetically closest $GPX6_{Cys}$ lineages, particularly the rabbit lineage, we further focus on the *Eumuroida* convergent sites. We use the inferred ancestral GPX sequences (**Section 5.3.2**) to identify 25 sites, excluding the Sec-to-Cys site, that change only over the *Eumuroida* branch. Of these sites, we identify 14 sites that show signatures of convergence across $GPX6_{Cys}$ lineages (to the exclusion of those identified from cat and walrus (Zhang and Kumar 1997)). We also infer a further 22 amino acid sites (19 substitutions and a 3 C-terminal extension) that changed between the end of the *Eumuroida* branch and the modern mouse GPX6 protein; m-$GPX6_{Cys+22}$. $CONVERG2$ was again used to identify which of these 19 substitutions demonstrate signatures of convergence across $GPX6_{Cys}$ lineages.

We test for enrichment of selection signatures (following the methodology outlined in **Section 5.3.5.2**) in the fifteen sites showing signatures of convergence along the *Eumuroida* lineage and the eight sites showing signatures of convergence on the branch just preceding the modern mouse protein.

5.3.5.5. Reconstructing Phylogenies According to Convergence

Using *PHYML* (Guindon et al. 2010), we reconstructed the mammalian tree given: a) the full GPX6 protein, b) the N-terminal of GPX6, c) the GPX domain of GPX6, d) the N-terminal of GPX6, e) the 26 sites that change across the *Eumuroida* branch (including the 14 sites that show changes across the *Eumuroida* branch and convergent changes across $GPx6_{Cys}$ branches.This was repeated for comparison using the full GPX3 and GPX5 proteins.

5.3.6. Assessing Catalytic Activity in Ancient and Modern Proteins

The following work was undertaken by collaborators: ***Qing Cheng*** *(Karolinska Institutet),* ***Elias SJ Arnér*** *(Karolinska Institutet, National Institute of Oncology),* ***Martin Floor*** *(Universitat de Vic - Universitat Central de Catalunya, Barcelona Supercomputing Center (BSC)),* ***Baldomero Oliva Miguel*** *(Universitat Pompeu Fabra),* ***Jordi Villà-Freixa*** *(Universitat de Vic - Universitat Central de Catalunya, Institut de Recerca i Innovació en Ciències de la Vida i de la Salut a la Catalunya Central (IRIS-CC)).*

5.3.6.1. Experimental Assessment of Catalytic Activity

Work undertaken by ***Qing Cheng*** *and* ***Elias SJ Arnér.***

The Eu-GPX6$_{Sec}$, Eu-GPX6$_{Cys}$ and Eu-GPX6$_{Cys+25}$ proteins were reconstructed from heterologous expression in *Escherichia coli*. A mutant *E. coli* strain that does not recognise UAG as a STOP codon was used, which results in a much higher yield of Eu-GPX6$_{Sec}$ than would otherwise be produced by *E. coli* with standard genetic code decoding. The catalytic activity of each protein, and the modern mouse protein, was evaluated by measuring the peroxidation activity on H_2O_2 with GSH.

5.3.6.2. Simulating Catalytic Activity

Work undertaken by ***Martin Floor****,* ***Baldomero Oliva Miguel*** *and* ***Jordi Villà-Freixa.***

Structures for the GPX6 orthologs and nodes of the ancestral sequence reconstructions were built using *AlphaFold*2 (Jumper et al. 2021). All protein sequences considered cysteines at their catalytic positions, given the inability to represent non-canonical residues for the "ab initio" model construction. We ran protein-ligand binding energy landscape explorations using the *PELE* software (Borrelli et al. 2005) for each protein structure, with ligands for the simulation being glutathione and glutathione disulfide.

Simulations were first run to discover catalytic poses with low global energies; the catalytic distance was considered as the closest sulphur-sulphur distance between the catalytic cysteine and the glutathione sulphur atoms. The lowest binding energy poses, filtered by a catalytic distance threshold below 4Å, were used to run a second *PELE* simulation, thus focusing on exploring this catalytic minimum binding energy configuration. Each simulation comprised 95 replicas of 100 equilibration steps that constrained the ligand to its starting position, followed by 1,000 *PELE* steps without any constraint over the ligand coordinates.

All simulation trajectories for the same ligand were simultaneously analysed using all ligand positions aligned to a common protein reference structure. A Time-structure

Independent Component Analysis (TICA) was built to find the common slowest-relaxing feature combination (Molgedey and Schuster 1994) with the *PyEMMA* library (Scherer et al. 2015). Finally, and separately for each protein and ligand simulation, the probabilities of visiting the slowest TICA coordinate (IC1) according to the catalytic distance (S-S) were plotted as a free energy map.

5.4. Results

5.4.1. Rate of Evolution Surrounding the Loss of Sec

We inferred five independent losses of Sec in GPX6$_{Sec}$ (**Fig. 5.1**, dashed green branches) across 22 mammals by reconstructing the ancestral sequence at each node of their phylogeny with *PAML* (Yang et al. 2005). These losses all occur in the last 64 million years (approximate times given in **Fig. 5.1**; (Huchon et al. 2002; Steppan et al. 2004; Hallström and Janke 2008; Chatterjee et al. 2009; Nyakatura and Bininda-Emonds 2012) and have resulted in multiple GPX6$_{Cys}$ lineages.

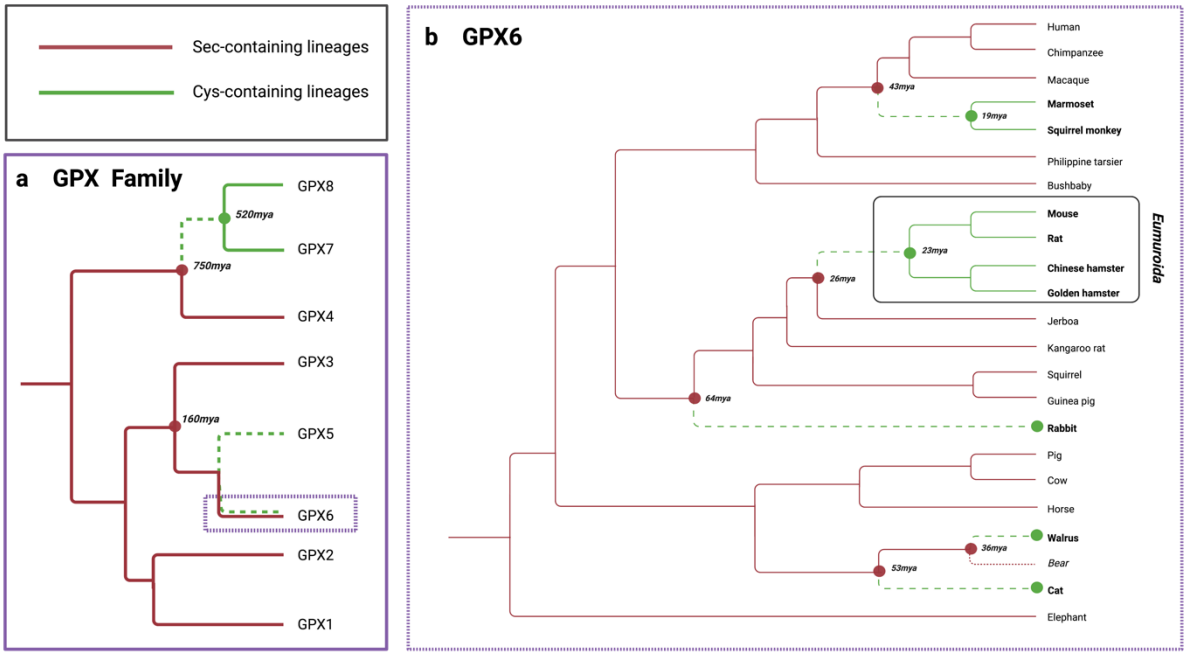


Fig. 5.1: Phylogenetic trees of the GPX family. *A) The phylogeny of the GPX family in Eukaryotes (based on (Mariotti et al. 2012)), including the dates of the duplications leading to GPX7$_{Cys}$, GPX8$_{Cys}$ and GPX5$_{Cys}$ and their older, single substitutions of Sec to Cys that resulted in enzymes with new properties. B) The topology of the phylogeny of the 22 mammals in our analysis. In red, GPX6$_{Sec}$ branches, in green, GPX6$_{Cys}$ ones. Dashed green branches represent GPX6$_{Cys}$ lineages where Sec was lost. Dotted red branch indicates the Bear GPX6$_{Sec}$ lineage, which was not used in the analysis due to sequence quality issues. The GPX6$_{Cys}$ Eumuroida clade, a specific group of muroid rodents, is boxed. Approximate ages given by (Huchon et al. 2002; Steppan et al. 2004; Higdon et al. 2007; Hallström and Janke 2008; Chatterjee et al. 2009; Nyakatura and Bininda-Emonds 2012).*

To measure the rate of evolution and approximate the degree of natural selection, we calculated independent dN/dS ratios for each branch in the mammalian tree, including ancestral branches (**Fig. 5.1**), and found higher dN/dS ratios in $GPX6_{Cys}$ lineages compared to neighbouring $GPX6_{Sec}$ lineages (**Fig. S5.1**). Hence, we infer faster evolution along the branches containing Cys in the place of Sec, which we then explicitly tested using the Branch model likelihood ratio test in *PAML* (Yang 2007).

When contrasting the dN/dS ratios of $GPX6_{Cys}$ lineages in the branches where Sec was lost (**Fig. 5.1**; dashed green branches) to $GPX6_{Cys}$ lineages in the branches inheriting this loss (**Fig. 5.1**, solid green branches) and $GPX6_{Sec}$ lineages (**Fig. 5.1**, solid red branches), we indeed observe a higher dN/dS ratio surrounding the times where Sec was substituted for Cys (LR test; P = 0.002; dN/dS = 0.370 dashed green versus 0.279 solid green versus 0.217 solid red branches in **Fig. 5.1**). Since our analysis excluded the Sec to Cys change, this suggests that a burst of amino acid evolution accompanied the loss of Sec.

5.4.1.1. Rate of Evolution Across Protein Domains

However, the higher dN/dS value observed across $GPX6_{Cys}$ lineages is still under 1, which is the threshold value often taken to confidently suggest positive selection is acting to increase the rate of evolution. Whilst the dN/dS ratio reaching this threshold value of 1 is unexpected in the case of otherwise strong constraint acting along a protein, particularly in the expected case of strongly conserved catalytic domains, it is possible that the inflated dN/dS value is instead a result of relaxed constraint, rather than positive selection necessarily acting surrounding the loss of Sec.

To further explore if the elevated dN/dS ratios were in line with the proposed positive selection, we repeated the previous likelihood ratio test over the three domains of the protein: the N-terminus, the GPX domain and the C-terminus (Mistry et al. 2021). Hence, we explicitly evaluate if the increased rate of evolution in the $GPX6_{Cys}$ lineages is focused to the protein's functional region. In the case of the GPX family of proteins, the GPX domain and, to a lesser extent, the C-terminus domain are essential for the activity of the enzyme and therefore noted as the functional regions. These domains both contain two key catalytic residues (U/C and Q in the GPX domain; W and N in the C-terminus domains) which together make the catalytic tetrad conserved across all $GPX6_{Sec}$ and $GPX6_{Cys}$ lineages (Toppo et al. 2008; Tosatto et al. 2008; Cheng and Arnér 2017). In contrast, the N-terminus is not thought essential to catalysis.

Reflecting this functional importance, we find that the GPX and N-terminus domains of $GPX6_{Sec}$ lineages to be most and least constrained, respectively, based on their dN/dS ratios (**Table 5.1**). However, the dN/dS ratio of the GPX domain is, unlike the N- and C-terminus, significantly larger in $GPX6_{Cys}$ lineages at the time Sec was lost (LR test; P = 2×10^{-5} ; where dN/dS = 0.384 in the dashed green, 0.186 in the solid green, 0.130 in the solid red branches in **Fig 5.1**). This is in further support of increased evolutionary change focused in the active GPX domain surrounding the time when Sec is abandoned in catalysis.

Table 5.1: dN/dS Ratios Calculated Across GPX proteins*. dN/dS ratios calculated for lineages where GPX6 has "Sec" (**Fig. 5.1**, solid red branches), has "Exchanged Sec for Cys" (**Fig. 5.1,** dashed green branches) or "Inherited Cys" (**Fig. 5.1**, solid green branches), and the number of identified convergent sites between lineages where GPX6 has gained Cys (**Fig. 5.1**, dashed green branches). dN/dS ratios and number of identified convergent sites for the GPX domain in other GPX proteins. The likelihood ratio test contrasts one ratio for all branches (null hypothesis) to different ratios among groups of branches. P-values are obtained from a χ^2 distribution with d.f = 2. $^{*}P < 0.05$; $^{**}P < 0.005$; $^{***}P < 0.0005$. In bold when significant and accompanied by sites under convergent evolution across $GPX6_{Cys}$ lineages.*

Protein	Region	Sec	Exchanged Sec for Cys	Inherited Cys	All	P-value	Convergent sites Number
$GPX6_{Cys}$	**Full length**	0.217	**0.370**	0.279	0.256	**0.002****	**22**
	N-terminus	0.436	0.411	0.671	0.460	0.268	3
	GPX domain	0.130	**0.384**	0.186	0.184	**2x10^{-5}*****	**12**
	C-terminus	0.174	0.258	0.250	0.203	0.157	7
$GPX1_{Sec}$	GPX domain	0.064	0.040	0.069	0.060	0.534	0
$GPX2_{Sec}$		0.075	0.042	0.038	0.060	0.191	0
$GPX3_{Sec}$		0.094	0.108	0.056	0.091	0.439	1
$GPX4_{Sec}$		0.062	0.007	0.203	0.061	$1x10^{-4}$***	0
$GPX5_{Sec}$		0.233	0.145	0.219	0.212	0.227	4
$GPX7_{Cys}$	GPX domain	0.083	0.080	0.117	0.088	0.712	0
$GPX8_{Cys}$		0.223	0.155	0.198	0.207	0.616	0

5.4.1.2. Rate of Evolution in the GPX Family

To validate that this observation is exclusive to $GPX6_{Cys}$, and therefore indicative of faster evolution associated with the Sec to Cys exchange rather than increased rate of evolution in the GPX domain due to its overall antioxidant function (Tian et al. 2021), we compared the rate of evolution in this domain to other enzymes in the GPX family.

We found no evidence of dN/dS inflation (**Table 5.1**, **Table S5.2**) across the GPX domain in the other GPX proteins for the analogous lineages where Sec was lost in GPX6 (analogous dashed green branches from **Fig 5.1**). Hence, neither Cys-containing GPX proteins nor the lineages where Cys is lost in GPX6 are otherwise inferred to display an inflated dN/dS value.

In $GPX4_{Sec}$, we do see a significant inflation in the dN/dS of the analogous lineages which have inherited Cys in GPX6 (analogous solid green branches from **Fig 5.1**) but we view

this as largely unrelated to the Sec to Cys exchange. Hence, we suggest that the dN/dS ratio of the GPX domain in $GPX6_{Cys}$ surrounding the time of the loss of Sec is unusually large for proteins of the GPX family.

5.4.2. Signatures of Adaptive Convergence

We now ask whether there is evidence for adaptive convergence on individual sites surrounding the time of Sec loss. We first use the Branch-Site test in *PAML* (Yang 2007) to ask if sites in the GPX domain show evidence for positive selection in the branches where Sec was lost (**Fig. 5.1**, dashed green branches) compared to all other branches. We see a significant enrichment of sites with such signatures (LR test; P = 0.046; **Table S5.3**), indicating the presence of sites in the GPX domain that show repeated changes (interpreted by this test as under positive selection) in $GPX6_{Cys}$ lineages where Sec was inferred to have been lost.

However, this Branch-Site test is non-significant when testing over the entire GPX6 protein. We reason that over more diverged lineages, epistasis limits the likelihood of the same sites showing repeated changes (Lunzer et al. 2010), and hence increases the probability of a false negative result of this test for positive selection. Indeed, we see that over the three most closely related lineages (branches leading to *Eumuroida,* rabbit and Cys-primate branch), this test results in a significant result over the entire protein (LR test; P = 0.008; **Table S5.3**). Most explicitly, this supports that changes of the same sites surround the loss of Sec in lineages with the most similar genetic backgrounds, likely due to the less differential role of epistasis over these lineages.

5.4.2.1. Convergent Sites Between Cys-branches

To more thoroughly explore which sites show repeated changes along the GPX6 phylogeny, we identify sites which show such convergent changes using $CONVERG2$ (Zhang and Kumar 1997). We see that convergence between lineages where Sec was lost (**Fig. 5.1**, dashed green branches) was the highest (**Table S5.4**), where the highest number of convergent sites are found in the GPX domain and the least in the N-terminus (54.6% in the GPX domain, followed by 31.8% and 13.6% in the C-terminus and N-terminus respectively). This approximately matches the lengths of each domain (113 sites of the GPX domain compared to the 65 and 39 sites of the C-terminus and N-terminus, respectively) despite the highest rate of amino acid changes being observed in the N-terminus (**Table 5.1**).

Moreover, convergence is largely subdued in the $GPX6_{Cys}$ lineages inheriting the loss of Sec (**Fig. 5.1**, solid green branches) and minimal in the $GPX6_{Sec}$ lineages, as well as for the other GPX proteins (**Tables S5.5-11**; **Fig. S5.2**). Further, simulations of protein evolution (modelled using $Seq - gen$ (Rambaut and Grassly 1997)), incorporating the accelerated rate of amino acid change in $GPX6_{Cys}$ sequences, fail to reproduce the pattern of convergence observed between these lineages at the time of Sec loss (**Fig. S5.3**). This remains true even when using the further accelerated rate of evolution as calculated in the GPX domain (**Fig. S5.4**). Further, these simulations also show that the few, weak convergence signatures in other GPX proteins are under expectations, based on their respective rate of amino acid change (**Figs. S5.5-11**). GPX3 and GPX5 are the most suitable GPX proteins to compare here owing to their more complete coding sequences, but we do stress that the overall pattern of convergence between the analogous lineages

to those containing Cys in GPX6 in all non-GPX6 proteins is that of much reduced convergence.

We observe that the highest level of convergence is between the basal *Eumuroida* (**Fig. 5.1**, dashed green line in box) and its genetically closer GPX6$_{Cys}$ lineages, particularly the rabbit (**Fig. S5.2, Table S5.4**). Of the 25 sites that change alongside the loss of Sec in the root of *Eumuroida* (**Fig. 5.2**, dashed green branch), 14 also show a site change in at least one of the other GPX6$_{Cys}$ lineages (**Fig. 5.2,** green box). These sites with convergent signatures are, again, mostly focused in the GPX catalytic domain (64.3%; **Table S5.4**) and enriched for signatures of positive selection that we observe along the branches leading to *Eumuroida*, rabbit and Cys-primates (as calculated by *PAML* (Yang 2007); M-W U test, P = $1.573e-7$), further supporting that these sites show an unusual degree of repeated change over these lineages. We also find an enrichment of signatures of positive selection, albeit weaker, in convergent sites in the GPX6$_{Cys}$ lineages following the loss of Sec (M-W U test, P=0.007) (**Fig. 5.2**, solid green branch),) but not preceding it, in agreement with adaptive convergence concentrated around the Sec to Cys exchange.

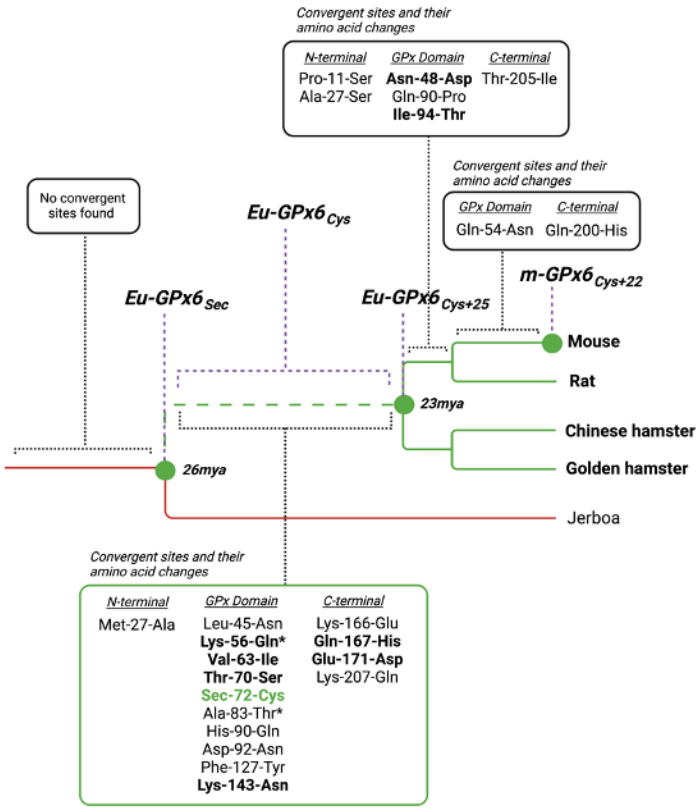


***Fig. 5.2: Topology of the phylogeny of the Eumuroida GPX6$_{Cys}$ clade**. Eumuroida GPX6$_{Cys}$ clade given as the green branches, with the Jerboa GPX6$_{Sec}$ lineage, red branch, as an outgroup. Amino acid exchanges showing signatures of convergence (CONVERG2; (Zhang and Kumar 1997)) across GPX6$_{Cys}$ lineages for each branch given in the boxes. Sites that have repeatedly changed in the GPX6$_{Cys}$ lineages towards similar or the same amino acid are shown in bold (green box).. Further, the * denotes sites with a posterior probability of positive selection in the upper 90th percentile across the GPX domain in GPX6$_{Cys}$ lineages, which are significantly enriched at the time Sec was abandoned.*

5.4.2.2. Phylogenetic Signatures of Convergence

Since adaptive convergence can mimic shared ancestry, it can often distort the topology of the species phylogeny (Edwards 2009). We find this to be the case here, with the tree reconstructed from the GPX domain (using *PHYML* (Guindon et al. 2010)) showing decreased divergence between the rabbit and *Eumuroida* clade (**Fig. S5.12D**). This is also observed, to a lesser extent, when reconstructing from the also catalytically relevant C-terminus (**Fig. S5.12E**), but not observed with the N-terminus domain (**Fig. S5.12C**) nor with other GPX proteins (**Fig. S5.13**).

If we reconstruct the mammalian phylogeny using the 15 convergent sites changing at the root of *Eumuroida* (**Fig. 5.3**, 14 identified in the dashed green box plus the Sec-to-Cys site) approximately 23-26 million years ago (Huchon et al. 2002), we see a striking departure from the species tree (**Fig. 5.3**). Despite their large divergences across the tree, the GPX6$_{Cys}$ species form two clades, as expected under the scenario of adaptive convergence. One clade is formed from the rabbit and *Eumuroida*, sharing a most recent common ancestry to the exclusion of all other species despite an approximately 64-million-year divergence (Hallström and Janke 2008). The remaining GPX6$_{Cys}$ lineages, which diverged approximately 100 million years ago (Hallström and Janke 2008), are then grouped also within a singular clade.

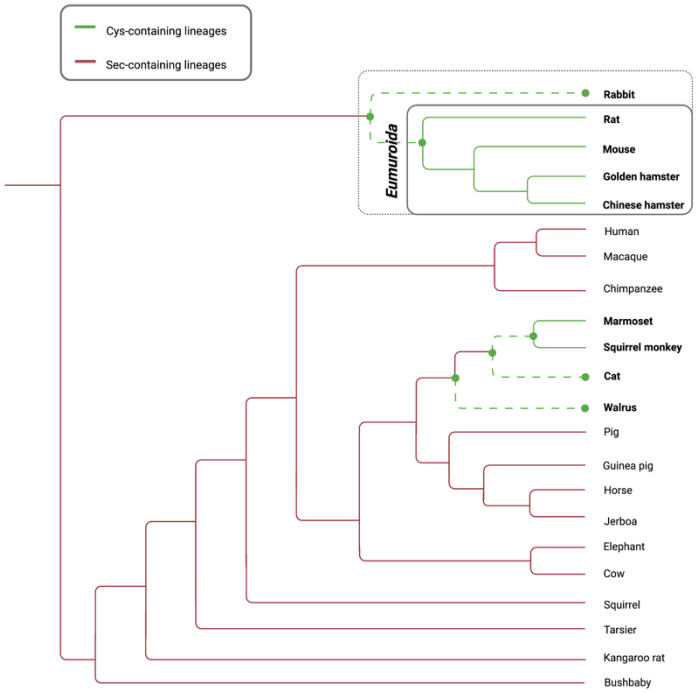


Figure 5.3: Phylogenetic tree of GPX6 reconstructed from convergent sites. *Topology of the phylogenetic tree, with midpoint rooting, from the 14 convergent sites accompanying the Sec to Cys substitution (**Fig. 5.2**, green box) in the basal Eumuroida GPX6$_{Cys}$ lineage (**Fig. 5.2**, dashed green branch). In sharp contrast to the species phylogeny (**Fig. 5.4.1B**), the GPX6$_{Sec}$ lineages now form two clades.*

5.4.3. Catalytic Activity of GPX Proteins

Given the observed signatures of adaptive convergence in the *Eumuroida,* we now focus on exploring the functional consequences of such changes in this lineage. We reconstructed three ancient proteins (**Fig. 5.2**), approximately dated to 23-26 million years ago, at the root of this clade and assessed them experimentally and computationally, alongside a fourth modern mouse protein.

These proteins are: 1) the ancestral protein before the loss of Sec, Eu-GPX6$_{Sec}$, taken from the common ancestor of the *Eumuroida* and Jerboa species 26 million years ago (Huchon et al. 2002); 2) the same ancestral protein with Cys instead of Sec, Eu-GPX6$_{Cys}$; 3) the ancestral but later-day protein with Cys and 25 other amino acids changes, Eu-GPX6$_{Cys+25}$, taken from the common ancestor of the *Eumuroida* species 23 million years ago (where 15 of 26 these amino acid changing sites, including the Cys site, have signatures of adaptive convergence; **Fig. 5.2**); and 4) the present-day mouse protein, m-GPX6$_{Cys+22}$, with 22 additional amino acid changes (19 substitutions and a 3 C-terminal extension) from Eu-GPX6$_{Cys+25}$ and no clear signatures of adaptive convergence (**Fig. 5.2**). In the latter protein, we also mutated the enzyme to contain either Sec or redox inactive serine (Ser) for comparisons of activity with the Sec- and Cys-variants.

The following work was carried out by our collaborators, as outlined in **Section 6.3.6.**

The reconstructed ancient and modern proteins were produced as recombinant proteins heterologously expressed in *Escherichia coli*. The Sec insertion system in bacteria is non-compatible with mammalian selenoprotein-encoding genes, hampering the production of proteins with Sec; thus, we employed a recently-developed method utilizing UAG redefined as a Sec codon in a release factor-1 deficient *E. coli* host strain lacking other UAG codons (Cheng and Arnér 2017).

We first compared the catalytic activity of Eu-GPX6$_{Sec}$ and Eu-GPX6$_{Cys}$ with H_2O_2 as the peroxide substrate and GSH as the reducing agent, with the expectation that substitution of Sec for Cys would lower its turnover (Axley et al. 1991; Berry et al. 1992; Johansson et al. 2005; Kim et al. 2015). Indeed, the ancient Eu-GPX6$_{Sec}$ protein displays the classic peroxidase activity of Sec-containing GPX enzymes, whereas Eu-GPX6$_{Cys}$, had almost no activity for this reaction (**Fig. 5.4A**).

The large drop in catalysis from Eu-GPX6$_{Sec}$ to Eu-GPX6$_{Cys}$ coincides with signatures of convergent adaptive evolution along the basal *Eumuroida* lineage (**Fig. 5.2**), initially suggesting a functional role of the accompanying amino acid changes to the loss of Sec. To ask if the additional 25 additional changes along the basal *Eumuroida* lineage recovered catalysis of this protein, we then measured the catalytic activity in Eu-GPX6$_{Cys+25}$ on H_2O_2 with GSH. Remarkably, classic GPX activity was not recovered (**Fig. 5.4B**). Finally, we turned to the extant m-GPX6$_{Cys+22}$ protein (**Fig. 5.4.2**), which is 90% identical to Eu-GPX6$_{Cys+25}$. Surprisingly, this Cys-containing variant also lacks classic GPX activity with H_2O_2 and GSH (**Fig. 5.4C**).

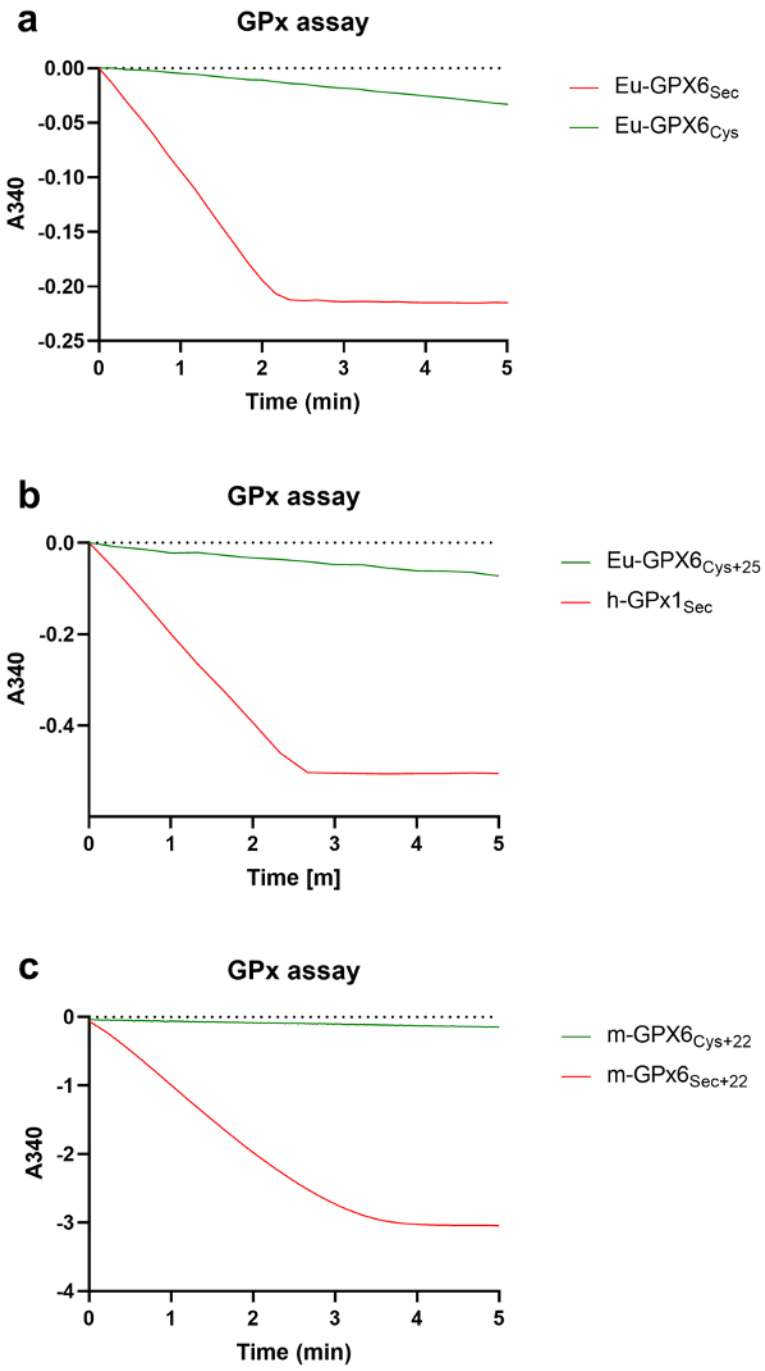


Fig. 5.4: Experimental assessment of catalytic function of GPX6 proteins*. (A) Experimental assessment of peroxidase reaction with H_2O_2 as a substrate for ancient Eu-GPX6$_{Sec}$ (red) and Eu-GPX6$_{Cys}$ (green). NADPH consumption by GR is indicated by the decrease in absorbance at 340 nm over time in the coupled assay (see* ***Section 5.3.6*** *for further details). (B) Equivalent assay for ancient Eu-GPX6$_{Cys+25}$ (green), which has very limited activity compared to human GPx1 (red) used here as a positive control. (C) Equivalent assay for modern m-GPX6$_{Cys+22}$ (green), again with scant activity, which is recovered once this protein is mutated to contain Sec, m-GPX6$_{Sec+22}$ (red).*

Together, this suggests that the inferred adaptive amino acid changes along this protein's evolution do not act to recapitulate Sec activity. However, this classic GPX activity is re-acquired when Cys is mutated back into Sec, producing the synthetic m-GPX6$_{Sec+22}$ variant (**Fig. 5.4C**). Indeed, our computational analysis suggests that the binding of GSH and overall structures of the enzymes (**Fig. 5.5A**) have not been adversely affected by the acquisition of Cys and that convergent amino acid substitutions are mainly located in the enzyme's surface (**Fig. 5.5B**). This is the case for the other GPX6$_{Cys}$ lineages, suggesting that all GPX6$_{Cys}$ mammals are able to recover classic GPX function with Sec (**Fig. S5.14**).

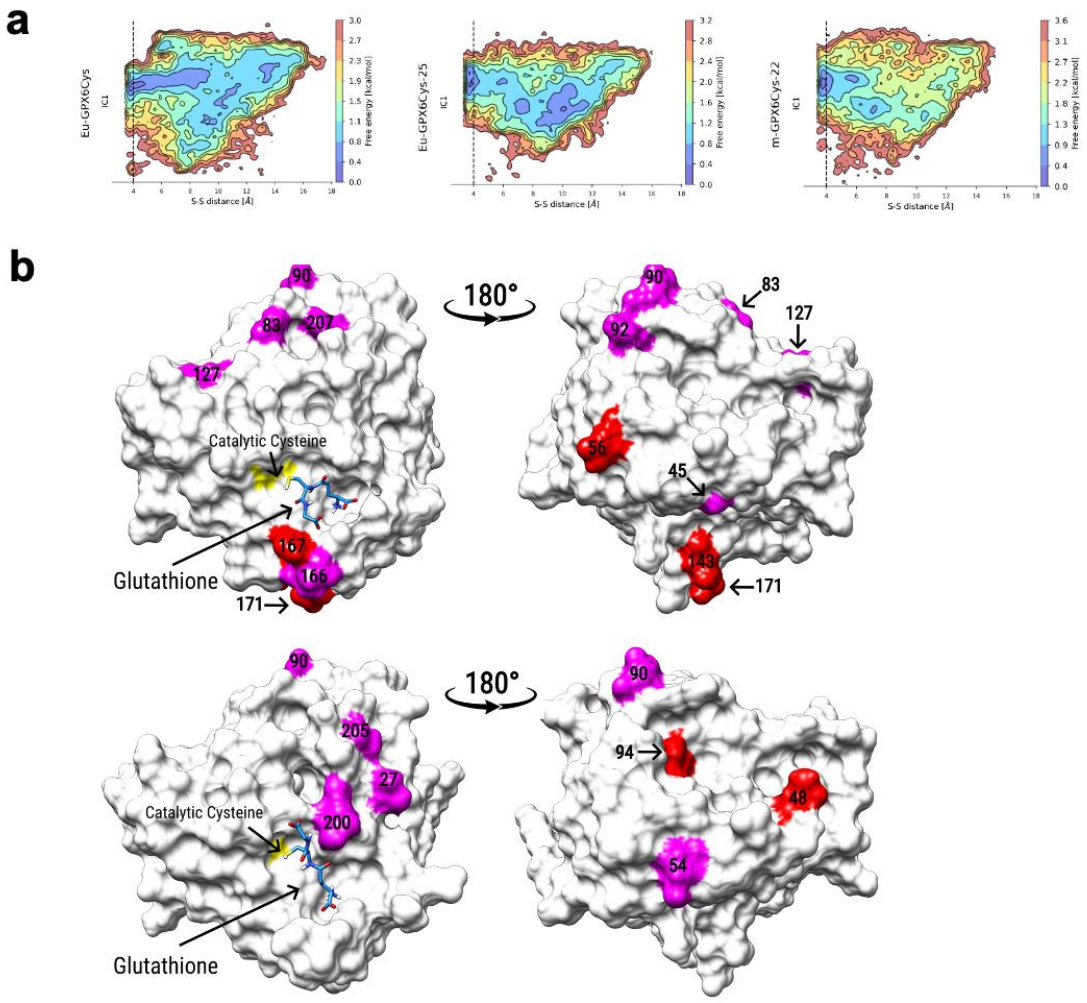


***Figure. 5.5*: Computational analysis of GPX6 catalytic function**. A). Free energy profiles for the docking of glutathione to Eu-GPX6Cys (left), Eu-GPX6Cys+25 (centre) and m-GPX6Cys+22 (right). The x-axis represents the distance between the catalytic cysteine sulphur atom and the ligand's sulphur atom, while the Y-axis shows the slowest TICA coordinate (IC1). The vertical dashed line represents a 4Å distance, with the free energy minimum in the three enzymes within this reactive catalytic distance. **B)** Convergence patterns (Fig 2a) from Eu-GPX6Cys to Eu-GPX6Cys+25 (top) and from Eu-GPX6Cys+25 to m-GPX6Cys+22 (Mouse-GPX6) (bottom). Sites converging towards similar (magenta) or the same (red) amino acids are shown with their sequence position. The catalytic cysteine (yellow) is shown with the glutathione best binding energy conformation (green) sampled during docking simulations.*

5.5. Discussion

The exchange of Sec for Cys in selenoproteins has long been linked to a reduction in catalytic activity, explained by the unique enzymatic properties of selenium in Sec. These include increased reactivity and nucleophilicity leading to improved catalysis (Arnér 2010), broader range of substates and pH in which catalysis is possible (Gromer et al. 2003) and perhaps increased resistance to oxidation (Snider et al. 2013). Whether these properties can always be reproduced by sulfur in Cys is unclear (Johansson et al. 2005) but the strong purifying selection on Sec sites in vertebrates (Castellano et al. 2009) suggests that not every reaction catalysed by Sec can be supported by Cys, at least not without some variation in these enzymatic properties.

Previous studies have shown that the catalytic activity conferred by Sec can be somewhat recovered with compensatory mutations following its exchange to Cys, but these may not restore full catalytic activity, with mutations in the Thioredoxin reductase of *Drosophila melanogaster* compensating no more than 50% of the catalytic rate in the human enzyme with Sec (Kanzok et al. 2001; Gromer et al. 2003). Still, Sec typically has low expression (Liu et al. 2012), possible due to inefficiencies in the Sec recoding process (Mehdi et al. 2013), and proteins with Cys may compensate in the way of increased expression rather than explicitly improving catalytic ability.

GPX6 presents a rare opportunity to investigate the evolutionary outcomes following the loss of the catalytically powerful Sec, and allows us to ask whether catalytic ability is lost, recovered or exchanged in modern Cys-containing orthologues. Here, we reconstruct ancient mammalian GPX6 proteins before and after the loss of Sec, and compare the evolutionary activity and mutational trajectories surrounding and following the exchange to Cys to that of lineages that have preserved Sec. We also experimentally reconstruct and assay the inferred ancient proteins in the *Eumuroida* lineage, alongside the experimental assay of the modern mouse protein, and present evidence towards the functional trajectory of this protein over its evolution.

5.5.1. Adaptive Convergence in $GPX6_{Cys}$

We demonstrate that substituting Sec for Cys in GPX6, and thereby abandoning selenium for sulfur in catalysis, leads to a burst of evolutionary activity in lineages sharing this exchange. These amino acid changes are not only concentrated in the functional domain but are often shared across $GPX6_{Cys}$ lineages, suggesting a narrow evolutionary exchange for GPX6 to recover functionality when losing Sec. This is likely limited by intragenic epistasis with linked sites, supported by increased convergence between our most closely related lineages, and the assumed preservation of other enzymatic properties expected to be important for overall activity (Fraïsse et al. 2019; Sharir-Ivry and Xia 2021). Importantly, such signatures of adaptive convergence are not observed in other GPX proteins.

5.5.2. Function of $GPX6_{Cys}$

However, we also show that whilst typical GPX activity is lost with the loss of Sec, it is not regained with the following mutations shared over Cys-lineages. This, especially considering the signatures of adaptive convergence, is unlikely to reflect non-functionality of modern $GPX6_{Cys}$ proteins, since the modern mouse protein has been shown to be expressed in the mouse embryo, testis, olfactory epithelium and brain

(Kryukov et al. 2003; Shema et al. 2015; Goltyaev et al. 2020) and when knocked down results in a deleterious neurological phenotype (Shema et al. 2015). We instead suggest that GPX6$_{Cys}$ proteins show evolutionary trajectories towards novel properties, as suggested to have occurred with GPX5$_{Cys}$, GPX7$_{Cys}$ and GPX8$_{Cys}$ enzymes that lost Sec much earlier (Herbette et al. 2007; Chen et al. 2016; Trenz et al. 2021).

Moreover, since the modern mouse GPX6$_{Cys}$ protein is able to recover classic GPX function with Sec (also computationally suggested for other mammalian GPX6$_{Cys}$ proteins), it is possible that its loss has resulted in subtly different enzymatic properties, whilst devoid of its classic function. This is also in agreement with what is known of the other Cys-containing GPXs, which act on alternative substrates for peroxidation (Nguyen et al. 2011; Taylor et al. 2013; Buday and Conrad 2021). Only comprehensive functional characterizations of these individual GPX6$_{Cys}$ enzymes in mammals will provide insights into the exact functional consequences of the observed convergent evolutionary trajectories, be that relating to peroxidation activity or otherwise. Further work may also aim to resolve the likely evolutionary order of shared amino acid changes across mammalian GPX6$_{Cys}$ lineages, providing insight into the epistatic interactions underlying such functional evolution.

5.5.3. Summary

In summary, we present the first evidence for molecular convergence of changes in proteins when abandoning unusual selenium in catalysis for common sulfur, hence ablating activity. These concerted changes follow a narrow path, maintaining some enzymatic properties and possibly adding new ones. Because multiple non-vertebrate species have completely abandoned enzymatic selenium for sulfur, we wonder whether other convergent adaptations leading to uncharted functions remain hidden in nature.

5.6. Conclusion

We reconstruct the evolutionary and functional trajectories of the mammalian GPX6 protein following its loss of its main catalytic residue selenocysteine. We show that signatures of adaptive convergence follow the exchange of selenocysteine for cysteine, but typical GPX catalytic activity is not recovered. Hence, we suggest that GPX6$_{Cys}$ proteins have gained yet unidentified abilities, acquired more recently, independently and convergently across lineages, instead of simply recovering the catalytic rate of their previous reaction.

Chapter 6: Discussion

6.1. Overview

In this thesis I have explored the role of micronutrients in genetic adaptation and in shaping genetic diversity in modern humans and wider mammals. I have considered the microevolution and macroevolution of genes associated with the uptake, metabolism and regulation of micronutrients, with particular attention to the adaptive, local response of human populations to micronutrient-associated selective pressures.

In **Chapter 1** I outlined the key theories relating to adaptive genome evolution. I then gave an overview of modern human evolutionary history, before summarising the current evidence for local adaptation across different modern human populations. I summarised the dominant methods used to identify local genetic adaptation, and presented the argument for micronutrient levels in the diet as a key selective driver in human populations. Throughout my discussions of human local adaptation, I referenced the issues of sampling biases and gaps amongst populations, particularly how this can contribute to differential health outcomes and clinical care. Finally, I described what is currently known on selenoprotein evolution, outlining the functional importance and evolution across vertebrates of the 21st amino acid selenocysteine.

In **Chapter 2,** I used a simulation framework to model local adaptation in four major human populations. I used this framework to test the power of different methods to identify the genomic signatures of positive selection on standing genetic variation, or "soft sweeps", at both the monogenic and polygenic level. I showed that the allele-differentiation statistic F_{ST} and recently developed genealogical method *Relate* have the highest power to identify local adaptation by selection on standard variation, including at times as recent as 10kya. Conversely, I report that, as expected, the haplotype-based methods often have low power to detect positive selection on standing variation on our simulated scenarios. These simulations are, to our knowledge, the most comprehensive evaluation to date of the power of the *Relate* method to identify the genomic signatures of positive selection, and I emphasise the promise of tree-recording methods in identifying more elusive signatures of positive selection. I also use these simulations to demonstrate the power of using an empirical neutral distribution to identify SNPs with signatures of positive selection according to *Relate*, and recommend this approach when using small sample sizes and when investigating the signatures of positive selection in multiple populations with different demographic histories.

I use these simulations to inform the methodology of **Chapters 3 and 4**. In **Chapter 3,** I use F_{ST} and *Relate* to investigate the signatures of natural selection in 40 diverse modern human populations across 276 micronutrient-associated genes for 13 micronutrients. This is the most comprehensive analysis of micronutrient-associated adaptation in modern humans to date, and allows the comparison of signatures of positive selection at multiple levels: amongst populations, amongst micronutrient categories and amongst individual micronutrient-associated genes. I show the presence of signatures of positive selection in genes associated with multiple different micronutrients, and across many geographic areas. I identify the populations and candidate genes with the strongest evidence for having undergone micronutrient-associated genetic adaptation, and show that some of these proposed adaptations are in

agreement with micronutrient soil levels and dietary deficiencies in contemporary populations. I find no evidence for classic polygenic models of positive selection, and instead infer that selection driven by micronutrient-associated selective pressures is more likely oligogenic than polygenic in nature. Ultimately, I propose that micronutrient levels in the diet are an important selective force in modern humans, and have contributed to the shaping of our genetic variation.

Chapter 4 is a natural extension to **Chapter 3,** where I focus on five micronutrients to discuss in detail the strength of evidence of positive selection, geographic breadth of putative positive selection, inferred polygenicity of the selective response, and relevance to contemporary human health issues. I identify the most likely genes which have mediated adaptive responses to micronutrient-associated pressures amongst populations, and identify the cases where the proposed genes differ between geographically separated populations. I use a combination of methods to explore the potential origin and timing of positive selection acting on these micronutrient-associated genes, and present evidence for a zinc-associated adaptation event in the Middle-East swiftly following the Out-of-Africa migration. I do identify a small number of cases where micronutrient-associated adaptation more likely occurred around the Neolithic, and propose that the selective drivers behind micronutrient-associated adaptation are likely not limited to soil composition.

Finally, in **Chapter 5** I explore the role of the micronutrient selenium in the evolution of a mammalian protein. We present the first evidence for molecular convergent evolution in proteins when exchanging selenocysteine for cysteine, inferring a narrow mutational path when losing the selenium-containing amino acid. Alongside suggested adaptive convergence, we show that there is also a loss of classic catalytic function when losing selenocysteine, and hypothesise that new enzymatic properties have been acquired by the GPX6 protein upon selenium loss. We propose the development of a novel functions across this selenoprotein, and further suggest the potential role of adaptive convergent evolution of non-vertebrate selenoproteins.

6.2. Local Adaptation in Modern Humans

Local adaptation has been inferred to have contributed to the modest genetic variation of our species. Identifying unknown instances of local adaptation in modern humans will thus contribute to a more comprehensive understanding of the evolutionary origin of human genetic diversity, particularly of diversity that contributes to population differences, which is particularly important when differentiated alleles may contribute to health inequality in contemporary populations. Here, I focus on, in my view, the most important open questions and future directions of the field, in light of the current literature and work developed in this thesis.

6.2.1. Selection on Standing Variation

The importance of selection on standing variation in modern humans has long been discussed (Hermisson and Pennings 2005, 2017; Prezeworski et al. 2005; Pritchard et al. 2010). However, with few recent exceptions (Schrider and Kern 2016, 2017), the confident inference of its role in human evolutionary history has been limited by the difficulty in identifying its individual genomic targets.

The low effective mutation rate in modern humans places a natural limitation on adaptation via *de novo* mutation (Hahn 2018). Segregating alleles, maintained by

balancing selection or drift, may more commonly mediate novel selective pressures, often encountered by humans when migrating to new environments. More than this, the extensive history of admixture in modern humans has allowed the frequent exchange of genetic variants that, whilst not truly novel, may be novel to the population receiving gene flow and thereby also facilitate local adaptation (but will not be addressed in detail here; (Ahlquist et al. 2021; Gopalan et al. 2022)). Hence, much of human local adaptation may be represented in the genome by subtle signatures of positive selection, particularly those driven by selection acting on alleles already segregating in a population, and therefore remain unidentified and undocumented. It must now be considered that the current bank of accepted or strongly supported examples of human local adaptation is considerably biased towards those conferred by *de novo* mutations (Pritchard et al. 2010; Schrider and Kern 2016; Rees et al. 2020) and, by extension, does not accurately represent the breadth of local adaptation in modern humans.

Thus, it is clear that the methods aiming to identify positive selection and consequent adaptation in human populations, especially with the aim of identifying new targets of positive selection, should be designed to consider SSV (not least because a wealth of methods already exist to identify selection on *de novo* mutation; (Weir and Cockerham 1984; Tajima 1989; Voight et al. 2006; Sabeti et al. 2007; Yi et al. 2010; Ferrer-Admetlla et al. 2014; Yassin et al. 2016; Crawford et al. 2017; Schmidt et al. 2019; Szpiech et al. 2021)). This likely include methods that, unlike classic summary statistics, do not simplify evolutionary history into a singular statistic value, but rather consider a more "full perspective". This may be in the way of integrating many patterns (including summary statistics) across loci, as in ABC or machine learning methods (Peter et al. 2012; Key et al. 2014, 2018; Pybus et al. 2015; Schrider and Kern 2016, 2017, 2018). Alternatively, this may be by considering the full history of a locus as in tree-recording methods (in reality, history as inferred up to the point of the common ancestor of sampled individuals; (Kelleher et al. 2019; Speidel et al. 2019; Hubisz and Siepel 2020b)). The key similarity in methods more suited to identifying SSV is their utilisation of the complexity of evolutionary patterns, which I suggest is the most appropriate avenue for identifying the weaker, and more variable, signatures of selection on standing variation.

Identifying SSV in human local adaptation likely also requires developments tangential to the field of genomics. Given the subtly of the signatures of SSV, and the increased difficulty in differentiating those signatures out from the neutral background of the genome, it is easy to imagine that the evidence for SSV is often weaker and less convincing. Thus, the importance of providing additional support to proposed selection on standing variation by other means, or additional orthogonal evidence, cannot be over-emphasised. This may be by identifying the same signatures of positive selection in different datasets or by use of different methods (although, some methods may be very closely related), or by functional assessment of the putatively adaptive variant, as addressed in the following section (**Section 6.2.3**). Perhaps the strongest supporting evidence for local adaptation is that of correlation of genomic signatures to proposed environmental factors, which independently may be viewed as evidence for positive selection (see **Section 6.2.2** for further discussion).

In terms of complex trait adaptation through polygenic adaptation, which may also be in-part driven by SSV, additional supporting evidence for adaptation may be given by the polygenic signatures of positive selection concentrated within a functional gene set or inferred directional change of a trait in a given population (Daub et al. 2013; Speidel

et al. 2019). However, the genetic architecture of complex traits may vary over populations (Mathieson 2021), in turn resulting in potentially different genes mediating adaptation amongst populations. Here, the importance of including diverse and under-studied populations in studies of genomic adaptation is further emphasised (addressed more fully in **Section 6.2.5**).

6.2.2. Environment as a Selective Driver

Correlations between signatures of positive selection and environmental factors can provide further support for claims of human local adaptation, given that the correlation is more extreme than what could be explained by shared ancestry (Fumagalli et al. 2011; Schlebusch et al. 2015; Key et al. 2018). For candidate genes demonstrating signatures of positive selection along a geographic cline, a correlation between such genomic signatures and an environmental factor along the same cline may indicate a likely selective pressure.

However, whilst some environmental factors are well-recorded across the globe (such as temperature, precipitation or UV, as available at www.worldclim.org), global documentation of other environmental factors relevant to human adaptation is often lacking. This includes soil composition and micronutrient content, as addressed in **Chapters 3 & 4**. In these cases, data (if even available) must often be integrated over different studies to provide a more global view of environmental variation, which is not always possible or reliable if data has come from studies of different designs or using different methods of data recording.

Moreover, the available data is often only at the resolution of the country or continental region, and does not represent fine-scale environmental variation. Local adaptation in modern humans to soils containing toxic levels of arsenic has been suggested in a singular region of Argentina (Schlebusch et al. 2015), and other such fine-scale adaptation to local environment, soil or otherwise, likely exist in other populations. However, without high-resolution environmental data across different environments, it is difficult to 1) form hypotheses of human local adaptation in response to environment; 2) contextualise identified signatures of positive selection; 3) validate signatures of positive selection (particularly important when signatures are more subtle). Therefore, limited environmental data can now be thought to actively restrict progress in our understanding the adaptive response of humans to their local environment.

A more comprehensive understanding of environmental factors throughout the globe may also identify where genetic variation, or genetic adaptations, are shared by populations experiencing the same environmental selective pressures. Adaptive convergence has previously been suggested in humans in response to high elevation (Yi et al. 2010; Bigham and Lee 2014; Huerta-Sánchez et al. 2014; Crawford et al. 2017) and in response to low-iodine soils in **Chapters 3 & 4.** Given how the same environmental selective pressures may be experienced across geographically disparate populations, and given the high degree of sharing of genetic variants amongst modern human populations, one can also hypothesise that other examples of convergent adaptation are likely to exist in humans.

6.2.3. Functional Evidence of Signatures of Positive Selection

Signatures of positive selection, however confidently identified, can be difficult to interpret without a clear adaptive function. Even if correlations exist between environmental factors and signatures of positive selection, it remains important to verify that putatively adaptive alleles are indeed driving an adaptive phenotype. Again, such functional information is especially important when considering weak signatures of positive selection (more so when environmental data is not available).

Given that many genes play a role in different functions, it is difficult to confidently link signatures of positive selection with the proposed adaptive phenotype without full functional assessment of the putatively selected allele. An overview of promising functional approaches is given in **Chapter 1** (**Section 1.6.6**) but includes integrating transcriptomics, metabolic and microbiotic datasets; high-throughput assays and potentially stem-cell technology (Kilpinen et al. 2017; Downes et al. 2019; Hwang et al. 2019; Zhou et al. 2022).

However, these functional approaches must also consider that the function of putatively adaptive alleles have the potential to differ amongst different genetic backgrounds. This is particularly pertinent when considering complex trait adaptation, where many alleles conferring adaptation may differ amongst different populations (Pritchard et al. 2010; Boyle et al. 2017; Mathieson 2021). Hence, I yet again emphasise the need for including more diverse cohorts of populations not only in studies of local adaptation, but also in those explicitly considering molecular function.

6.2.4. Importance of Studies over Diverse Populations

In recent years, there has been an explosion of genomic data of modern humans, including that from ancient DNA (Racimo et al. 2015; Wohns et al. 2022). Despite this, there is still a clear bias towards certain populations in studies of human genetic diversity, particularly towards Europeans in GWAS (Sirugo et al. 2019). This imbalanced representation of human populations has no doubt led to a biased representation of the genetic diversity of modern humans and the failure to capture a non-trivial portion of the genetic variation within our species. This makes the aim of understanding the origin and function of the genetic diversity of modern humans impossible, and does not allow a full fully informed evaluation of the contribution of genetic variation to population phenotypic differences, including those of traits relevant to health. In the dawn of personalised and genomic medicine, the failure to document and understand the genetic variation of all populations has the potential to substantially contribute to contemporary health inequalities (outlined in **Chapter 1**; **Section 1.4.3**).

Indeed, multiple well-supported examples of local adaptation in modern humans have direct contemporary health consequences (Kwiatkowski 2005; Genovese et al. 2010; Wang et al. 2012; Clemente et al. 2014; Mathieson et al. 2015a; Minster et al. 2016; Key et al. 2018). This includes adaptations that 1) increase the risk of various metabolic disorders in a contemporary environment; 2) cause various inherited disorders; 3) result in differences in the efficacy of treatment for non-inherited disorders (see **Section 1.4.2** for further details and discussion).

More thorough population sampling may allow the identification of other examples of local adaptation pertinent to modern health, potentially mediated by novel adaptive variants in currently undocumented or poorly sampled populations. Moreover, increased sample size of poorly sampled populations will increase the power of population genetics methods to identify variants mediating adaptation, and will further facilitate an understanding of 1) the role of local selective drivers in human adaptation; 2) the genomic response to selective pressures; and 3) the relationship between genetic variation and adaptive phenotype (particularly important when such an adaptive phenotype can affect disease risk or progression).

Finally, the inclusion of under-represented populations in GWAS is particularly vital in providing a deeper understanding of polygenic adaptation and the genetic architecture of complex phenotypes. Complex traits are expected to be driven by variants that may differ amongst populations (Pritchard et al. 2010; Mathieson 2021) but many of the alleles inferred to carry a risk for a complex disease phenotype are currently inferred through Euro-centric GWAS and therefore cannot be expected to be replicated across diverse populations (Sirugo et al. 2019). Thus, we currently lack reliable estimates of risk for many complex diseases amongst many populations which, if used in clinical care, can result in poorly informed medical decisions and increased health risk. It is clear that this disproportionally affects under-studied populations (Sirugo et al. 2019) and is the strongest incentive to including more diverse populations in any study exploring complex traits or polygenic adaptation.

6.2.5. Summary

Progress in the field of local adaptation, in my view, relies on three main factors. The first is the development of methodology integrating entire (or more complete) evolutionary patterns to identify more subtle signatures of positive selection, including selection on standing variation. This will improve our understanding of human genetic variation and identify currently undocumented or only putative cases of local adaptation. The second is the integration of complimentary data, such as environment and functional data, which will supply the necessary support for currently undocumented or only putative examples of local adaptation. The final is the most important factor, and can be considered the rate-limiting step across all aspects of local adaptation progress: increased sampling of undocumented and understudied populations. This is not only necessary to understand the genetic diversity of our species, including that driven by local adaptation, but is fundamental in understanding the relationship between genetic variation and medically-relevant traits amongst global populations.

6.3. Selenium in Macroevolution

Selenium has long been known to play a key catalytic role in selenoproteins, being the defining element of their constituent amino acid, selenocysteine (Chambers et al. 1986; Stadtman 1996). The evolutionary constraint acting on the selenocysteine in selenoproteins, and the consequent low exchangeability between selenocysteine and cysteine in these proteins, has also suggested selenium as an important element throughout vertebrate evolution (Castellano et al. 2009), Indeed, in **Chapter 5**, we show that losing selenium may drive a mammalian protein to develop currently unknown functions, rather than performing the same functions available to a protein containing selenium.

This then drives a collection of key questions pertaining to the function of this Cys-containing protein and selenoproteome diversity, and how that may be related to the loss of a selenium-containing amino acid. If there is indeed a completely novel function in GPX6 following the loss of Sec, we can also ask: do other selenoproteins compensate for the loss of classic activity? Alternatively, if the Cys-containing GPX6 is able to continue peroxidation by acting on different substrates as suggested in other selenoproteins losing selenocysteine (Nguyen et al. 2011; Taylor et al. 2013; Buday and Conrad 2021), one can ask: do the losses of Sec in those proteins drive the same level of convergence as observed in GPX6 (as described in **Chapter 5**)? This is only a small set of open questions, but the answers to these (and others) rely on 1) further functional assessment of Cys-containing selenoproteins and 2) a greater understanding of the genetic diversity of the other selenoproteins in mammalian lineages where selenoproteins have been lost or maintained.

Understanding the role of selenium in macroevolution should also include non-mammalian taxa. It has been suggested that the higher levels of selenium in the aquatic environment of fish has increased their dependence on this rare element (Sarangi et al. 2017), ultimately resulting in their large selenoproteome (Castellano et al. 2009; Mariotti et al. 2012). It can be expected that evolutionary pathways and functional consequences following the loss of selenocysteine (and selenium) may differ from that of mammals: the larger dependence on environmental selenium may have locked this taxon into maintaining selenoprotein function, and less novel functions may evolve in comparison to mammals. Alternatively, the larger selenoproteome may result in more successful functional compensation by selenoproteins maintaining Sec, and the development of new functions may be less constrained than inferred in mammals. Again, understanding the loss or gain of protein function following the loss of selenium requires both extensive functional analysis and an understanding of the genetic diversity of the entire selenoproteome of taxa, rather than individual selenoproteins.

6.4. Thesis Conclusion

In this thesis, I have explored the role micronutrients have played in both micro and macroevolution, particularly in driving local adaptation in modern humans. Ultimately, I present work that demonstrates the importance of considering micronutrients in the evolution of our species and across wider biology. Finally, I highlight the recent developments that present the most promise in furthering our understanding of human local adaptation, and outline the key limiting factors.

References

Adams, F. and Hathcock, P. J., 1984. Aluminum Toxicity and Calcium Deficiency in Acid Subsoil Horizons of Two Coastal Plains Soil Series. *Soil Science Society of America Journal*, 48 (6), 1305–1309.

Adhikari. K., Reales, G., Smith, A.J., Konka, E, Palmen, J., Quinto-Sanchez, M., Acuña-Alonzo, V., Jaramillo, C., Arias, W., Fuentes, M., . . . Ruiz-Linares, A. 2015 A genome-wide association study identifies multiple loci for variation in human ear morphology. *Nat Commun*. 24;6:7500

Ahlquist, K. D., Bañuelos, M. M., Funk, A., Lai, J., Rong, S., Villanea, F. A. and Witt, K. E., 2021. Our Tangled Family Tree: New Genomic Methods Offer Insight into the Legacy of Archaic Admixture. *Genome Biology and Evolution*, 13 (7), evab115.

Ajib, F. A. and Childress, J. M., 2022. *Magnesium Toxicity* [online]. StatPearls [Internet]. StatPearls Publishing. Available from: https://www.ncbi.nlm.nih.gov/books/NBK554593/ [Accessed 19 Mar 2023].

Akey, J. M., Eberle, M. A., Rieder, M. J., Carlson, C. S., Shriver, M. D., Nickerson, D. A. and Kruglyak, L., 2004. Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes. *PLOS Biology*, 2 (10), e286.

Akey, J. M., Swanson, W. J., Madeoy, J., Eberle, M. and Shriver, M. D., 2006. TRPV6 exhibits unusual patterns of polymorphism and divergence in worldwide populations. *Human Molecular Genetics*, 15 (13), 2106–2113.

Al Alawi, A. M., Majoni, S. W. and Falhammar, H., 2018. Magnesium and Human Health: Perspectives and Research Directions. *International Journal of Endocrinology*, 2018, 1–17.

Alewell, C., Ringeval, B., Ballabio, C., Robinson, D. A., Panagos, P. and Borrelli, P., 2020. Global phosphorus shortage will be aggravated by soil erosion. *Nature Communications*, 11 (1), 4546.

Alexander, D. H., Novembre, J. and Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19 (9), 1655–1664.

Allentoft, M. E., Sikora, M., Sjögren, K.-G., Rasmussen, S., Rasmussen, M., Stenderup, J., Damgaard, P. B., Schroeder, H., Ahlström, T., . . . Willerslev, E., 2015. Population genomics of Bronze Age Eurasia. *Nature*, 522 (7555), 167–172.

Alloway, B. J., 2013. Heavy Metals and Metalloids as Micronutrients for Plants and Animals. *In*: Alloway, B. J., ed. *Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability* [online]. Dordrecht: Springer Netherlands, 195–209. Available from: https://doi.org/10.1007/978-94-007-4470-7_7 [Accessed 13 Mar 2023].

Alloway, B. J. and Tills, A. R., 1984. Copper deficiency in world crops. *Outlook on Agriculture*, 13 (1), 32–42.

Amorim, C. E. G., Daub, J. T., Salzano, F. M., Foll, M. and Excoffier, L., 2015. Detection of Convergent Genome-Wide Signals of Adaptation to Tropical Forests in Humans. *PLOS ONE*, 10 (4), e0121557.

Anand, L. and Rodriguez Lopez, C. M., 2022. ChromoMap: an R package for interactive visualization of multi-omics data and annotation of chromosomes. *BMC Bioinformatics*, 23 (1), 33.

Anderson, J. J. B. and Allen, J. C., 1994. Nutrition of Macrominerals and Trace Elements. *In*: Goldberg, I., ed. *Functional Foods: Designer Foods, Pharmafoods, Nutraceuticals* [online]. Boston, MA: Springer US, 323–354. Available from: https://doi.org/10.1007/978-1-4615-2073-3_15 [Accessed 30 Jan 2023].

Andrés, A. M., 2011. Balancing Selection in the Human Genome. *In*: *eLS* [online]. John Wiley & Sons, Ltd. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0022863 [Accessed 28 Feb 2023].

Archer, N. M., Petersen, N., Clark, M. A., Buckee, C. O., Childs, L. M. and Duraisingh, M. T., 2018. Resistance to Plasmodium falciparum in sickle cell trait erythrocytes is driven by oxygen-dependent growth inhibition. *Proceedings of the National Academy of Sciences*, 115 (28), 7350–7355.

Armitage, S. J., Jasim, S. A., Marks, A. E., Parker, A. G., Usik, V. I. and Uerpmann, H.-P., 2011. The Southern Route "Out of Africa": Evidence for an Early Expansion of Modern Humans into Arabia. *Science*, 331 (6016), 453–456.

Arnér, E. S. J., 2010. Selenoproteins—What unique properties can arise with selenocysteine in place of cysteine? *Experimental Cell Research*, 316 (8), 1296–1303.

Arnér, E. S. J. and Holmgren, A., 2006. The thioredoxin system in cancer. *Seminars in Cancer Biology*, 16 (6), 420–426.

Arunachalam, P., Kannan, P., Prabukumar, G. and Govindaraj, M., n.d. Zinc deficiency in Indian soils with special focus to enrich zinc in peanut.

Ashish, B., Neeti, K. and Himanshu, K., 2013. Copper Toxicity: A Comprehensive Study, 2.

Aspray, T. J., Yan, L. and Prentice, A., 2005. Parathyroid hormone and rates of bone formation are raised in perimenopausal rural Gambian women. *Bone*, 36 (4), 710–720.

Asthana, S., Schmidt, S., Sunyaev., S. 2005. A limited role for balancing selection, *Trends in Genetics,* 21(1), 30-32.

Axley, M. J., Böck, A. and Stadtman, T. C., 1991. Catalytic properties of an Escherichia coli formate dehydrogenase mutant in which sulfur replaces selenium. *Proceedings of the National Academy of Sciences of the United States of America*, 88 (19), 8450–8454.

Bai, H., Guo, X., Zhang, D., Narisu, N., Bu, J., Jirimutu, J., Liang, F., Zhao, X., Xing, Y., . . . Zhou, H., 2014. The Genome of a Mongolian Individual Reveals the Genetic Imprints of Mongolians on Modern Human Populations. *Genome Biology and Evolution*, 6 (12), 3122–3136.

Bailey, R. L., Jr, K. P. W. and Black, R. E., 2015. The Epidemiology of Global Micronutrient Deficiencies. *Annals of Nutrition and Metabolism*, 66 (Suppl. 2), 22–33.

Barceloux, D. G. and Barceloux, D., 1999. Molybdenum. *Journal of Toxicology: Clinical Toxicology*, 37 (2), 231–237.

Barroso, G. V., Puzović, N. and Dutheil, J. Y., 2019. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15 (11), e1008449.

Baumann, P., Lee, J., Frossard, E., Schönholzer, L. P., Diby, L., Hgaza, V. K., Kiba, D. I., Sila, A., Sheperd, K. and Six, J., 2021. Estimation of soil properties with mid-infrared soil spectroscopy across yam production landscapes in West Africa. *SOIL*, 7 (2), 717–731.

Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., . . . Kelleher, J., 2022. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220 (3), iyab229.

Becker, M. and Asch, F., 2005. Iron toxicity in rice—conditions and management concepts. *Journal of Plant Nutrition and Soil Science*, 168 (4), 558–573.

Beltrame, M. H., Rubel, M. A. and Tishkoff, S. A., 2016. Inferences of African evolutionary history from genomic data. *Current Opinion in Genetics & Development*, 41, 159–166.

Berg, J. J. and Coop, G., 2014. A Population Genetic Signal of Polygenic Adaptation. *PLOS Genetics*, 10 (8), e1004412.

Berg, J. J., Harpak, A., Sinnott-Armstrong, N., Joergensen, A. M., Mostafavi, H., Field, Y., Boyle, E. A., Zhang, X., Racimo, F., Pritchard, J. K. and Coop, G., 2019. Reduced signal for polygenic adaptation of height in UK Biobank. *eLife*, 8, e39725.

Berg, J. J., Zhang, X. and Coop, G., 2019. Polygenic Adaptation has Impacted Multiple Anthropometric Traits. [online]. Available from: https://www.biorxiv.org/content/10.1101/167551v4 [Accessed 26 Jan 2023].

Bergström, A., McCarthy, S. A., Hui, R., Almarri, M. A., Ayub, Q., Danecek, P., Chen, Y., Felkel, S., Hallast, P., Kamm, J., Blanché, H., Deleuze, J.-F., Cann, H., Mallick, S., Reich, D., Sandhu, M. S., Skoglund, P., Scally, A., Xue, Y., Durbin, R. and Tyler-Smith, C., 2020. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367 (6484), eaay5012.

Berislav, M., 1999. A Case Report of Acute Human Molybdenum Toxicity from a Dietary Molybdenum Supplement – A New Member of the »Lucor Metallicum« Family. *Arhiv za higijenu rada i toksikologiju*, 50 (3), 289–297.

Berry, M. J., Maia, A. L., Kieffer, J. D., Harney, J. W. and Larsen, P. R., 1992. Substitution of cysteine for selenocysteine in type I iodothyronine deiodinase reduces the catalytic efficiency of the protein but enhances its translation. *Endocrinology*, 131 (4), 1848–1852.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D,E., Hirschhorn, J.N. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet*. 74(6):1111-20

Beyer, R. M., Krapp, M., Eriksson, A. and Manica, A., 2021. Climatic windows for human migration out of Africa in the past 300,000 years. *Nature Communications*, 12 (1), 4889.

Bhatia, G., Patterson, N., Sankararaman, S. and Price, A. L., 2013. Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23 (9), 1514–1521.

Bhutta, Z. A. and Salam, R. A., 2012. Global Nutrition Epidemiology and Trends. *Annals of Nutrition and Metabolism*, 61 (Suppl. 1), 19–27.

Biban, B. and Lichiardopol, C., 2017. Iodine Deficiency, Still a Global Problem? *Current Health Sciences Journal*, 43 (2), 103.

Bigham, A. W. and Lee, F. S., 2014. Human high-altitude adaptation: forward genetics meets the HIF pathway. *Genes & Development*, 28 (20), 2189–2204.

Bitarello, B. D., de Filippo, C., Teixeira, J. C., Schmidt, J. M., Kleinert, P., Meyer, D. and Andrés, A. M., 2018. Signatures of Long-Term Balancing Selection in Human Genomes. *Genome Biology and Evolution*, 10 (3), 939–955.

Blome, M. W., Cohen, A. S., Tryon, C. A., Brooks, A. S. and Russell, J., 2012. The environmental context for the origins of modern human diversity: a synthesis of regional variability in African climate 150,000-30,000 years ago. *Journal of Human Evolution*, 62 (5), 563–592.

Borrelli, K. W., Vitalis, A., Alcantara, R. and Guallar, V., 2005. PELE: Protein Energy Landscape Exploration. A Novel Monte Carlo Based Technique. *Journal of Chemical Theory and Computation*, 1 (6), 1304–1311.

Bowler, J. M., Johnston, H., Olley, J. M., Prescott, J. R., Roberts, R. G., Shawcross, W. and Spooner, N. A., 2003a. New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature*, 421 (6925), 837–840.

Bowler, J. M., Johnston, H., Olley, J. M., Prescott, J. R., Roberts, R. G., Shawcross, W. and Spooner, N. A., 2003b. New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature*, 421 (6925), 837–840.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A.

G. and Bustamante, C. D., 2008. Assessing the Evolutionary Impact of Amino Acid Mutations in the Human Genome. *PLOS Genetics*, 4 (5), e1000083.
Boyle, E. A., Li, Y. I. and Pritchard, J. K., 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169 (7), 1177–1186.
Brajesh, R. G., Dutta, D. and Saini, S., 2019. Distribution of fitness effects of mutations obtained from a simple genetic regulatory network model. *Scientific Reports*, 9 (1), 9842.
Brandt, D., Wei, X., Deng, Y., Vaughn, A. H., Nielsen, R., 2022. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics.* 221(1).
Brissot, P., Pietrangelo, A., Adams, P. C., de Graaff, B., McLaren, C. E. and Loréal, O., 2018. Haemochromatosis. *Nature reviews. Disease primers*, 4, 18016.
Brown, T. A., Jones, M. K., Powell, W. and Allaby, R. G., 2009. The complex origins of domesticated crops in the Fertile Crescent. *Trends in Ecology & Evolution*, 24 (2), 103–109.
Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S. and Akey, J. M., 2018. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell*, 173 (1), 53-61.e9.
Buday, K. and Conrad, M., 2021. Emerging roles for non-selenium containing ER-resident glutathione peroxidases in cell signaling and disease. *Biological Chemistry*, 402 (3), 271–287.
Buffalo, V. and Coop, G., 2020. Estimating the genome-wide contribution of selection to temporal allele frequency change. *Proceedings of the National Academy of Sciences*, 117 (34), 20672–20680.
Burger, J., Link, V., Blöcher, J., Schulz, A., Sell, C., Pochon, Z., Diekmann, Y., Žegarac, A., Hofmanová, Z., . . . Wegmann, D., 2020. Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years. *Current biology: CB*, 30 (21), 4307-4315.e13.
Buxbaum, J., Jacobson, D. R., Tagoe, C., Alexander, A., Kitzman, D. W., Greenberg, B., Thaneemit-Chen, S. and Lavori, P., 2006. Transthyretin V122I in African Americans With Congestive Heart Failure. *Journal of the American College of Cardiology*, 47 (8), 1724–1725.
Caballero, B., 2002a. Global Patterns of Child Health: The Role of Nutrition. *Annals of Nutrition and Metabolism*, 46 (Suppl. 1), 3–7.
Caballero, B., 2002b. Global Patterns of Child Health: The Role of Nutrition. *Annals of Nutrition and Metabolism*, 46 (Suppl. 1), 3–7.
Caldas, I. V., Clark, A. G., Messer, P. V. 2022. Inference of selective sweep parameters through supervised learning. *bioRxiv*. doi: https://doi.org/10.1101/2022.07.19.500702
Calcium, I. of M. (US) C. to R. D. R. I. for V. D. and, Ross, A. C., Taylor, C. L., Yaktine, A. L. and Valle, H. B. D., 2011. *Tolerable Upper Intake Levels: Calcium and Vitamin D* [online]. Dietary Reference Intakes for Calcium and Vitamin D. National Academies Press (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK56058/ [Accessed 19 Mar 2023].
Campbell, A. D., Colombatti, R., Andemariam, B., Strunk, C., Tartaglione, I., Piccone, C. M., Manwani, D., Asare, E. V., Boruchov, D., . . . Antwi-Boasiako, C., 2021. An Analysis of Racial and Ethnic Backgrounds within the CASiRe International Cohort of Sickle Cell Disease Patients: Implications for Disease Phenotype and Clinical Research. *Journal of racial and ethnic health disparities*, 8 (1), 99–106.

Campbell, M. C. and Tishkoff, S. A., 2008. AFRICAN GENETIC DIVERSITY: Implications for Human Demographic History, Modern Human Origins, and Complex Disease Mapping. *Annual review of genomics and human genetics*, 9, 403–433.

Cao, L., Zhang, W., Liu, X., Yang, P., Wang, J., Hu, K., Zhang, X., Liu, W., He, X., Jing, H. and Yuan, X., 2019. The Prognostic Significance of PDE7B in Cytogenetically Normal Acute Myeloid Leukemia. *Scientific Reports*, 9, 16991.

Carlberg, C., 2022. Vitamin D in the Context of Evolution. *Nutrients*, 14 (15), 3018.

Carter, P. and Wells, J. A., 1988. Dissecting the catalytic triad of a serine protease. *Nature*, 332 (6164), 564–568.

Castellano, S., Andrés, A. M., Bosch, E., Bayes, M., Guigó, R. and Clark, A. G., 2009. Low Exchangeability of Selenocysteine, the 21st Amino Acid, in Vertebrate Proteins. *Molecular Biology and Evolution*, 26 (9), 2031–2040.

Castellano, S., Lobanov, A. V., Chapple, C., Novoselov, S. V., Albrecht, M., Hua, D., Lescure, A., Lengauer, T., Krol, A., Gladyshev, V. N. and Guigó, R., 2005. Diversity and functional plasticity of eukaryotic selenoproteins: Identification and characterization of the SelJ family. *Proceedings of the National Academy of Sciences*, 102 (45), 16188–16193.

Castellano, S., Morozova, N., Morey, M., Berry, M. J., Serras, F., Corominas, M. and Guigó, R., 2001. In silico identification of novel selenoproteins in the Drosophila melanogaster genome. *EMBO reports*, 2 (8), 697–702.

Castellano, S., Parra, G., Sánchez-Quinto, F. A., Racimo, F., Kuhlwilm, M., Kircher, M., Sawyer, S., Fu, Q., Heinze, A., . . . Pääbo, S., 2014. Patterns of coding variation in the complete exomes of three Neandertals. *Proceedings of the National Academy of Sciences*, 111 (18), 6666–6671.

Castiglioni, S., Cazzaniga, A., Albisetti, W. and Maier, J. A. M., 2013. Magnesium and Osteoporosis: Current State of Knowledge and Future Research Directions. *Nutrients*, 5 (8), 3022–3033.

Cha, H. J., Jang, D. S., Kim, Y.-G., Hong, B. H., Woo, J.-S., Kim, K.-T. and Choi, K. Y., 2013. Rescue of Deleterious Mutations by the Compensatory Y30F Mutation in Ketosteroid Isomerase. *Molecules and Cells*, 36 (1), 39–46.

Chambers, I., Frampton, J., Goldfarb, P., Affara, N., McBain, W. and Harrison, P. R., 1986. The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon, TGA. *The EMBO Journal*, 5 (6), 1221–1227.

Chang, A. R. and Anderson, C., 2017. Dietary Phosphorus Intake and the Kidney. *Annual Review of Nutrition*, 37, 321–346.

Charlesworth, B., Morgan, M. T. and Charlesworth, D., 1993. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134 (4), 1289–1303.

Charlesworth, D., Charlesworth, B. and Morgan, M. T., 1995. The pattern of neutral molecular variation under the background selection model. *Genetics*, 141 (4), 1619–1632.

Chatterjee, H. J., Ho, S. Y., Barnes, I. and Groves, C., 2009. Estimating the phylogeny and divergence times of primates using a supermatrix approach. *BMC Evolutionary Biology*, 9 (1), 259.

Chen, Y. and Barak, P., 1982. Iron Nutrition of Plants in Calcareous Soils. *In*: Brady, N. C., ed. *Advances in Agronomy* [online]. Academic Press, 217–240. Available from: https://www.sciencedirect.com/science/article/pii/S0065211308603260 [Accessed 16 Mar 2023].

Chen, Y.-I., Wei, P.-C., Hsu, J.-L., Su, F.-Y. and Lee, W.-H., 2016. NPGPx (GPx7): a novel oxidative stress sensor/transmitter with multiple roles in redox homeostasis. *American Journal of Translational Research*, 8 (4), 1626–1640.

Cheng, Q. and Arnér, E. S. J., 2017. Selenocysteine Insertion at a Predefined UAG Codon in a Release Factor 1 (RF1)-depleted Escherichia coli Host Strain Bypasses Species Barriers in Recombinant Selenoprotein Translation. *The Journal of Biological Chemistry*, 292 (13), 5476–5487.

Chevin, L.-M. and Hospital, F., 2008. Selective Sweep at a Quantitative Trait Locus in the Presence of Background Genetic Variation. *Genetics*, 180 (3), 1645–1660.

Cifor, 2006. *Forests and human health* [online]. Center for International Forestry Research (CIFOR). Available from: http://www.cifor.org/library/2088/forests-and-human-health/ [Accessed 14 Feb 2023].

Clark, J. D., Beyene, Y., WoldeGabriel, G., Hart, W. K., Renne, P. R., Gilbert, H., Defleur, A., Suwa, G., Katoh, S., Ludwig, K. R., Boisserie, J.-R., Asfaw, B. and White, T. D., 2003. Stratigraphic, chronological and behavioural contexts of Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature*, 423 (6941), 747–752.

Clemente, F. J., Cardona, A., Inchley, C. E., Peter, B. M., Jacobs, G., Pagani, L., Lawson, D. J., Antão, T., Vicente, M., . . . Kivisild, T., 2014. A Selective Sweep on a Deleterious Mutation in CPT1A in Arctic Populations. *The American Journal of Human Genetics*, 95 (5), 584–589.

Conrad, M., Jakupoglu, C., Moreno, S. G., Lippl, S., Banjac, A., Schneider, M., Beck, H., Hatzopoulos, A. K., Just, U., Sinowatz, F., Schmahl, W., Chien, K. R., Wurst, W., Bornkamm, G. W. and Brielmeier, M., 2004. Essential Role for Mitochondrial Thioredoxin Reductase in Hematopoiesis, Heart Development, and Heart Function. *Molecular and Cellular Biology*, 24 (21), 9414–9423.

Cornelis, M. C., Fornage, M., Foy, M., Xun, P., Gladyshev, V. N., Morris, S., Chasman, D. I., Hu, F. B., Rimm, E. B., Kraft, P., Jordan, J. M., Mozaffarian, D. and He, K., 2015. Genome-wide association study of selenium concentrations. *Human Molecular Genetics*, 24 (5), 1469–1477.

Costas, J., 2018. The highly pleiotropic gene SLC39A8 as an opportunity to gain insight into the molecular pathogenesis of schizophrenia. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, 177 (2), 274–283.

Covert, A. W., Lenski, R. E., Wilke, C. O. and Ofria, C., 2013. Experiments on the role of deleterious mutations as stepping stones in adaptive evolution. *Proceedings of the National Academy of Sciences*, 110 (34), E3171–E3178.

Crawford, J. E., Amaru, R., Song, J., Julian, C. G., Racimo, F., Cheng, J. Y., Guo, X., Yao, J., Ambale-Venkatesh, B., . . . Nielsen, R., 2017. Natural Selection on Genes Related to Cardiovascular Health in High-Altitude Adapted Andeans. *The American Journal of Human Genetics*, 101 (5), 752–767.

Cruciani, F., Trombetta, B., Massaia, A., Destro-Bisol, G., Sellitto, D. and Scozzari, R., 2011. A Revised Root for the Human Y Chromosomal Phylogenetic Tree: The Origin of Patrilineal Diversity in Africa. *American Journal of Human Genetics*, 88 (6), 814.

Cruickshank, T. E. and Hahn, M. W., 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23 (13), 3133–3157.

Cui, J., Pan, Y.-H., Zhang, Y., Jones, G. and Zhang, S., 2011. Progressive Pseudogenization: Vitamin C Synthesis and Its Loss in Bats. *Molecular Biology and Evolution*, 28 (2), 1025–1031.

Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R. M., Clegg, J. B., Langaney, A. and Excoffier, L., 2002. Molecular Analysis of the β-Globin Gene Cluster in the Niokholo

Mandenka Population Reveals a Recent Origin of the βS Senegal Mutation. *The American Journal of Human Genetics*, 70 (1), 207–223.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., and 1000 Genomes Project Analysis Group, 2011. The variant call format and VCFtools. *Bioinformatics*, 27 (15), 2156–2158.

Dannemann, M., Andrés, A. M. and Kelso, J., 2016. Introgression of Neandertal- and Denisovan-like Haplotypes Contributes to Adaptive Variation in Human Toll-like Receptors. *The American Journal of Human Genetics*, 98 (1), 22–33.

Darwin, C., 1859. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. London: John Murray.

Darwin, C. and Wallace, A., 1858. On the Tendency of Species to form Varieties; and on the Perpetuation of Varieties and Species by Natural Means of Selection. *Journal of the Proceedings of the Linnean Society of London. Zoology*, 3 (9), 45–62.

Daub, J. T., Dupanloup, I., Robinson-Rechavi, M. and Excoffier, L., 2015. Inference of Evolutionary Forces Acting on Human Biological Pathways. *Genome Biology and Evolution*, 7 (6), 1546–1558.

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M. and Excoffier, L., 2013. Evidence for polygenic adaptation to pathogens in the human genome. *Molecular Biology and Evolution*, 30 (7), 1544–1558.

Daub, J. T., Moretti, S., Davydov, I. I., Excoffier, L. and Robinson-Rechavi, M., 2017. Detection of Pathways Affected by Positive Selection in Primate Lineages Ancestral to Humans. *Molecular Biology and Evolution*, 34 (6), 1391–1402.

Davis, B. H., Poon, A. F. Y. and Whitlock, M. C., 2009. Compensatory mutations are repeatable and clustered within proteins. *Proceedings of the Royal Society B: Biological Sciences*, 276 (1663), 1823–1827.

De Groote, H., Tessema, M., Gameda, S. and Gunaratna, N. S., 2021. Soil zinc, serum zinc, and the potential for agronomic biofortification to reduce human zinc deficiency in Ethiopia. *Scientific Reports*, 11 (1), 8770.

Dean, A. M. and Golding, G. B., 1997. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. *Proceedings of the National Academy of Sciences*, 94 (7), 3104–3109.

Delaneau, O., Zagury, J.-F. and Marchini, J., 2013. Improved whole-chromosome phasing for disease and population genetic studies. *Nature Methods*, 10 (1), 5–6.

Demény, A., Kern, Z., Czuppon, G., Németh, A., Schöll-Barna, G., Siklósy, Z., Leél-Őssy, S., Cook, G., Serlegi, G., . . . Bondár, M., 2019. Middle Bronze Age humidity and temperature variations, and societal changes in East-Central Europe. *Quaternary International*, 504, 80–95.

Dhaliwal, S. S., Naresh, R. K., Mandal, A., Singh, R. and Dhaliwal, M. K., 2019. Dynamics and transformations of micronutrients in agricultural soils as influenced by organic matter build-up: A review. *Environmental and Sustainability Indicators*, 1–2, 100007.

Diallo, A. B., Makarenkov, V. and Blanchette, M., 2010. Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*, 26 (1), 130–131.

Diamond, J., 2002. Evolution, consequences and future of plant and animal domestication. *Nature*, 418 (6898), 700–707.

Dib, M.-J., Elliott, R. and Ahmadi, K. R., 2019. A critical evaluation of results from genome-wide association studies of micronutrient status and their utility in the practice of precision nutrition. *British Journal of Nutrition*, 122 (2), 121–130.

Distante, S., Robson, K. J. H., Graham-Campbell, J., Arnaiz-Villena, A., Brissot, P. and Worwood, M., 2004. The origin and spread of the HFE-C282Y haemochromatosis mutation. *Human Genetics*, 115 (4), 269–279.

Dobrovolskaya, M. V., 2005. Upper Palaeolithic and Late Stone Age Human Diet. *Journal of PHYSIOLOGICAL ANTHROPOLOGY and Applied Human Science*, 24 (4), 433–438.

Domínguez-Andrés, J., Kuijpers, Y., Bakker, O. B., Jaeger, M., Xu, C.-J., Van der Meer, J. W., Jakobsson, M., Bertranpetit, J., Joosten, L. A., Li, Y. and Netea, M. G., 2021. Evolution of cytokine production capacity in ancient and modern European populations. *eLife*, 10, e64971.

Dormitzer, P. R., Ellison, P. T. and Bode, H. H., 1989. Anomalously low endemic goiter prevalence among Efe pygmies. *American Journal of Physical Anthropology*, 78 (4), 527–531.

Drouin, G., Godin, J.-R. and Pagé, B., 2011. The Genetics of Vitamin C Loss in Vertebrates. *Current Genomics*, 12 (5), 371–378.

Duborská, E., Šebesta, M., Matulová, M., Zvěřina, O. and Urík, M., 2022. Current Strategies for Selenium and Iodine Biofortification in Crop Plants. *Nutrients*, 14 (22), 4717.

Durbin, R., Eddy, S. R., Krogh, A. and Mitchison, G., 1998. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* [online]. 1st ed. Cambridge University Press. Available from: https://www.cambridge.org/core/product/identifier/9780511790492/type/book [Accessed 11 Jan 2023].

Duret, L., 2008. Neutral Theory: The Null Hypothesis of Molecular Evolution. *Nature Education*, 1 (1), 803.

Durvasula, A. and Sankararaman, S., 2020. Recovering signals of ghost archaic introgression in African populations. *Science Advances*, 6 (7), eaax5097.

Dusseldor, G., Lombard, M. and Wurz, S., 2013. Pleistocene Homo and the updated Stone Age sequence of South Africa. *South African Journal of Science*, 109 (5–6), 01–07.

Edmeades, D., Morton, J., Waller, J., Metherell, A., Roberts, A. and Carey, P., 2010. The diagnosis and correction of potassium deficiency in New Zealand pastoral soils: a review. *New Zealand Journal of Agricultural Research*, 53 (2), 151–173.

Edwards, S. V., 2009. Natural selection and phylogenetic analysis. *Proceedings of the National Academy of Sciences*, 106 (22), 8799–8800.

Enard, D., Cai, L., Gwennap, C. and Petrov, D. A., n.d. Viruses are a dominant driver of protein adaptation in mammals. *eLife*, 5, e12469.

Enard, D. and Petrov, D. A., 2018. Evidence that RNA Viruses Drove Adaptive Introgression between Neanderthals and Modern Humans. *Cell*, 175 (2), 360-371.e13.

Enard, W., 2016. The Molecular Basis of Human Brain Evolution. *Current Biology*, 26 (20), R1109–R1117.

Engelken, J., Carnero-Montoro, E., Pybus, M., Andrews, G. K., Lalueza-Fox, C., Comas, D., Sekler, I., Rasilla, M. de la, Rosas, A., . . . Bosch, E., 2014. Extreme Population Differences in the Human Zinc Transporter ZIP4 (SLC39A4) Are Explained by Positive Selection in Sub-Saharan Africa. *PLOS Genetics*, 10 (2), e1004128.

Engelken, J., Espadas, G., Mancuso, F. M., Bonet, N., Scherr, A.-L., Jímenez-Álvarez, V., Codina-Solà, M., Medina-Stacey, D., Spataro, N., . . . Bosch, E., 2016. Signatures of Evolutionary Adaptation in Quantitative Trait Loci Influencing Trace Element Homeostasis in Liver. *Molecular Biology and Evolution*, 33 (3), 738–754.

Erdman, J. W., MacDonald, I. A. and Zeisel, S. H., 2012. *Present Knowledge in Nutrition: Tenth Edition* [online]. Wiley-Blackwell. Available from:

http://www.scopus.com/inward/record.url?scp=84878020186&partnerID=8YFLogxK [Accessed 19 Mar 2023].

Esoh, K. and Wonkam, A., 2021. Evolutionary history of sickle-cell mutation: implications for global genetic medicine. *Human Molecular Genetics*, 30 (R1), R119–R128.

Evershed, R. P., Davey Smith, G., Roffet-Salque, M., Timpson, A., Diekmann, Y., Lyon, M. S., Cramp, L. J. E., Casanova, E., Smyth, J., . . .Thomas, M. G., 2022. Dairying, diseases and the evolution of lactase persistence in Europe. *Nature*, 608 (7922), 336–345.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C. and Foll, M., 2013. Robust Demographic Inference from Genomic and SNP Data. *PLOS Genetics*, 9 (10), e1003905.

Fadhlaoui-Zid, K., Haber, M., Martínez-Cruz, B., Zalloua, P., Benammar Elgaaied, A. and Comas, D., 2013. Genome-Wide and Paternal Diversity Reveal a Recent Origin of Human Populations in North Africa. *PLoS ONE*, 8 (11), e80293.

Fagny, M., Patin, E., Macisaac, J. L., Rotival, M., Flutre, T., Jones, M. J., Siddle, K. J., Quach, H., Harmant, C., . . . Quintana-Murci, L., 2015. The epigenomic landscape of African rainforest hunter-gatherers and farmers. *Nature Communications*, 6.

Fagundes, N. J.R., Tagliani-Ribeiro, A., Rubicz, R., Tarskaia, L., Crawford, M.H., Salzano, F.M., Bonatto, S.L. 2018. How strong was the bottleneck associated to the peopling of the Americas? New insights from multilocus sequence data. *Genet Mol Biol*. 41(1):206-214.

Fan, S., Hansen, M. E. B., Lo, Y. and Tishkoff, S. A., 2016. Going global by adapting local: A review of recent human adaptation. *Science*, 354 (6308), 54–59.

Fan, S., Spence, J. P., Feng, Y., Hansen, M. E. B., Terhorst, J., Beltrame, M. H., Ranciaro, A., Hirbo, J., Beggs, W., . . .Tishkoff, S. A., 2023. Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell*, 186 (5), 923-939.e14.

Fernandes, V., Alshamali, F., Alves, M., Costa, M. D., Pereira, J. B., Silva, N. M., Cherni, L., Harich, N., Cerny, V., Soares, P., Richards, M. B. and Pereira, L., 2012. The Arabian Cradle: Mitochondrial Relicts of the First Steps along the Southern Route out of Africa. *American Journal of Human Genetics*, 90 (2), 347–355.

Fernando, D. R. and Lynch, J. P., 2015. Manganese phytotoxicity: new light on an old problem. *Annals of Botany*, 116 (3), 313–319.

Ferrer-Admetlla, A., Liang, M., Korneliussen, T. and Nielsen, R., 2014. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution*, 31 (5), 1275–1291.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I. and Pritchard, J. K., 2016a. Detection of human adaptation during the past 2000 years. *Science (New York, N.Y.)*, 354 (6313), 760–764.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I. and Pritchard, J. K., 2016b. Detection of human adaptation during the past 2000 years. *Science*, 354 (6313), 760–764.

de Filippo, C., Key, F. M., Ghirotto, S., Benazzo, A., Meneu, J. R., Weihmann, A., NISC Comparative Sequence Program, Parra, G., Green, E. D. and Andrés, A. M., 2016. Recent Selection Changes in Human Genes under Long-Term Balancing Selection. *Molecular Biology and Evolution*, 33 (6), 1435–1447.

FISHER, G., 2008. Micronutrients and Animal Nutrition and the Link between the Application of Micronutrients to Crops and Animal Health. *Turkish Journal of Agriculture and Forestry*, 32 (3), 221–233.

Fisher, R. A., 1919. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth and Environmental Science Transactions of The Royal Society of Edinburgh*, 52 (2), 399–433.

Fishilevich, S., Nudel, R., Rappaport, N., Hadar, R., Plaschkes, I., Iny Stein, T., Rosen, N., Kohn, A., Twik, M., Safran, M., Lancet, D. and Cohen, D., 2017. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database*, 2017, bax028.

Foll, M., Gaggiotti, O. E., Daub, J. T., Vatsiou, A. and Excoffier, L., 2014. Widespread Signals of Convergent Adaptation to High Altitude in Asia and America. *American Journal of Human Genetics*, 95 (4), 394–407.

Fraga, C. G., 2005. Relevance, essentiality and toxicity of trace elements in human health. *Molecular Aspects of Medicine*, 26 (4), 235–244.

Fraga, C. G. and Oteiza, P. I., 2002. Iron toxicity and antioxidant nutrients. *Toxicology*, 180 (1), 23–32.

Fraïsse, C., Puixeu Sala, G. and Vicoso, B., 2019. Pleiotropy Modulates the Efficacy of Selection in Drosophila melanogaster. *Molecular Biology and Evolution*, 36 (3), 500–515.

Freitas, S. R. S., 2018. Molecular Genetics of Salt-Sensitivity and Hypertension: Role of Renal Epithelial Sodium Channel Genes. *American Journal of Hypertension*, 31 (2), 172–174.

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., Johnson, P. L. F., Aximu-Petri, A., Prüfer, K., de Filippo, C., . . . Pääbo, S., 2014. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 514 (7523), 445–449.

Fujimoto, A., Kimura, R., Ohashi, J., Omi, K., Yuliwulandari, R., Batubara, L., Mustofa, M. S., Samakkarn, U., Settheetham-Ishida, W., Ishida, T., Morishita, Y., Furusawa, T., Nakazawa, M., Ohtsuka, R. and Tokunaga, K., 2008. A scan for genetic determinants of human hair morphology: EDAR is associated with Asian hair thickness. *Human Molecular Genetics*, 17 (6), 835–843.

Fumagalli, M., Moltke, I., Grarup, N., Racimo, F., Bjerregaard, P., Jørgensen, M. E., Korneliussen, T. S., Gerbault, P., Skotte, L., . . . Nielsen, R., 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science*, 349 (6254), 1343–1347.

Fumagalli, M., Sironi, M., Pozzoli, U., Ferrer-Admettla, A., Pattini, L. and Nielsen, R., 2011. Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure through Human Evolution. *PLOS Genetics*, 7 (11), e1002355.

Gagnon, K. B. and Delpire, E., 2013. Physiology of SLC12 transporters: lessons from inherited human genetic mutations and genetically engineered mouse knockouts. *American Journal of Physiology-Cell Physiology*, 304 (8), C693–C714.

Garcia, R. S., 2004. The Misuse of Race in Medical Diagnosis. *Pediatrics*, 113 (5), 1394–1395.

Garud, N. R., Messer, P. W., Buzbas, E. O. and Petrov, D. A., 2015. Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLOS Genetics*, 11 (2), e1005004.

Gebremichael, G., Demena, M., Egata, G. and Gebremichael, B., 2020. Prevalence of Goiter and Associated Factors Among Adolescents in Gazgibla District, Northeast Ethiopia. *Global Advances in Health and Medicine*, 9, 2164956120923624.

Geerling, J. C. and Loewy, A. D., 2008. Central regulation of sodium appetite. *Experimental Physiology*, 93 (2), 177–209.

Gerbault, P., Liebert, A., Itan, Y., Powell, A., Currat, M., Burger, J., Swallow, D. M. and Thomas, M. G., 2011. Evolution of lactase persistence: an example of human niche construction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366 (1566), 863–877.

Germeshausen, M., Ancliff, P., Estrada, J., Metzler, M., Ponstingl, E., Rütschle, H., Schwabe, D., Scott, R. H., Unal, S., Wawer, A., Zeller, B. and Ballmaier, M., 2018. MECOM-associated syndrome: a heterogeneous inherited bone marrow failure syndrome with amegakaryocytic thrombocytopenia. *Blood Advances*, 2 (6), 586–596.

Gibson, R. S., 2012. A Historical Review of Progress in the Assessment of Dietary Zinc Intake as an Indicator of Population Zinc Status123. *Advances in Nutrition*, 3 (6), 772–782.

Giri, A., Ranjan, P. and Bharti, V. K., 2021. Selenium Toxicity in Domestic Animals. *In*: *Selenium Contamination in Water* [online]. John Wiley & Sons, Ltd, 51–72. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119693567.ch4 [Accessed 14 Mar 2023].

Gloux, A., Le Roy, N., Brionne, A., Bonin, E., Juanchich, A., Benzoni, G., Piketty, M.-L., Prié, D., Nys, Y., Gautron, J., Narcy, A. and Duclos, M. J., 2019. Candidate genes of the transcellular and paracellular calcium absorption pathways in the small intestine of laying hens. *Poultry Science*, 98 (11), 6005–6018.

Gokhman, D., Malul, A. and Carmel, L., 2017. Inferring Past Environments from Ancient Epigenomes. *Molecular Biology and Evolution*, 34 (10), 2429–2438.

Goltyaev, M. V., Mal'tseva, V. N. and Varlamova, E. G., 2020. Expression of ER-resident selenoproteins and activation of cancer cells apoptosis mechanisms under ER-stress conditions caused by methylseleninic acid. *Gene*, 755, 144884.

González, A. M., Larruga, J. M., Abu-Amero, K. K., Shi, Y., Pestano, J. and Cabrera, V. M., 2007. Mitochondrial lineage M1 traces an early human backflow to Africa. *BMC Genomics*, 8 (1), 223.

Goodman, M. and Sterner, K. N., 2010. Phylogenomic evidence of adaptive evolution in the ancestry of humans. *Proceedings of the National Academy of Sciences*, 107 (supplement_2), 8918–8923.

Gopalan, S., Smith, S. P., Korunes, K., Hamid, I., Ramachandran, S. and Goldberg, A., 2022. Human genetic admixture through the lens of population genomics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377 (1852), 20200410.

Gouy, A., Daub, J. T. and Excoffier, L., 2017. Detecting gene subnetworks under selection in biological pathways. *Nucleic Acids Research*, 45 (16), e149.

Gouy, A. and Excoffier, L., 2020. Polygenic Patterns of Adaptive Introgression in Modern Humans Are Mainly Shaped by Response to Pathogens. *Molecular Biology and Evolution*, 37 (5), 1420–1433.

Gower, G., Picazo, P. I., Fumagalli, M., Racimo, F. 2021. Detecting adaptive introgression in human evolution using convolutional neural networks. *Elife*. 10.

Gravel, S., Henn, B. M., Gutenkunst, R. N., Indap, A. R., Marth, G. T., Clark, A. G., Yu, F., Gibbs, R. A., The 1000 Genomes Project, Bustamante, C. D., . . .McVean, G. A., 2011. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108 (29), 11983–11988.

Gravel, S., Zakharia, F., Moreno-Estrada, A., Byrnes, J. K., Muzzio, M., Rodriguez-Flores, J. L., Kenny, E. E., Gignoux, C. R., Maples, B. K., . . . Bustamante, C. D., 2013. Reconstructing Native American Migrations from Whole-Genome and Whole-Exome Data. *PLOS Genetics*, 9 (12), e1004023.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M. H.-Y., . . . Pääbo, S., 2010. A Draft Sequence of the Neandertal Genome. *Science (New York, N.Y.)*, 328 (5979), 710–722.

Greenhouse, R., 1981. Preparation effects on iron and calcium in traditional Pima foods. *Ecology of Food and Nutrition*, 10 (4), 221–225.

Greger, J. L., 1999. Nutrition versus toxicology of manganese in humans: evaluation of potential biomarkers. *Neurotoxicology*, 20 (2–3), 205–212.

Grillo, A., Salvi, L., Coruzzi, P., Salvi, P. and Parati, G., 2019. Sodium Intake and Hypertension. *Nutrients*, 11 (9), 1970.

Gromer, S., Johansson, L., Bauer, H., Arscott, L. D., Rauch, S., Ballou, D. P., Williams, C. H., Schirmer, R. H. and Arnér, E. S. J., 2003. Active sites of thioredoxin reductases: Why selenoproteins? *Proceedings of the National Academy of Sciences*, 100 (22), 12618–12623.

Grossman, H., Duggan, E., McCamman, S., Welchert, E. and Hellerstein, S., 1980. The Dietary Chloride Deficiency Syndrome. *Pediatrics*, 66 (3), 366–374.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., Zuk, O., Lander, E. S., Schaffner, S. F. and Sabeti, P. C., 2010. A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection. *Science*, 327 (5967), 883–886.

Grün, R., Stringer, C., McDermott, F., Nathan, R., Porat, N., Robertson, S., Taylor, L., Mortimer, G., Eggins, S. and McCulloch, M., 2005. U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *Journal of Human Evolution*, 49 (3), 316–334.

Gruss, L. T. and Schmitt, D., 2015. The evolution of the human pelvis: changing adaptations to bipedalism, obstetrics and thermoregulation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370 (1663), 20140063.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O., 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59 (3), 307–321.

Günther, T. and Coop, G., 2013. Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, 195 (1), 205–220.

Haak, W., Balanovsky, O., Sanchez, J. J., Koshel, S., Zaporozhchenko, V., Adler, C. J., Der Sarkissian, C. S. I., Brandt, G., Schwarz, C., . . . the Genographic Consortium, 2010. Ancient DNA from European Early Neolithic Farmers Reveals Their Near Eastern Affinities. *PLoS Biology*, 8 (11), e1000536.

Haber, M., Jones, A. L., Connell, B. A., Asan, Arciero, E., Yang, H., Thomas, M. G., Xue, Y. and Tyler-Smith, C., 2019. A Rare Deep-Rooting D0 African Y-Chromosomal Haplogroup and Its Implications for the Expansion of Modern Humans Out of Africa. *Genetics*, 212 (4), 1421–1428.

Hahn, M., 2018. *Molecular Population Genetics*. Oxford, New York: Oxford University Press.

Haldane, J. B. S., 1924. A Mathematical Theory of Natural and Artificial Selection. Part Ii the Influence of Partial Self-Fertilisation, Inbreeding, Assortative Mating, and Selective Fertilisation on the Composition of Mendelian Populations, and on Natural Selection. *Biological Reviews*, 1 (3), 158–163.

Haller, B. C. and Messer, P. W., 2019. SLiM 3: Forward Genetic Simulations Beyond the Wright–Fisher Model. *Molecular Biology and Evolution*, 36 (3), 632–637.

Hallström, B. M. and Janke, A., 2008. Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. *BMC Evolutionary Biology*, 8 (1), 162.

Halsted, J. A., Ronaghy, H. A., Abadi, P., Haghshenass, M., Amirhakemi, G. H., Barakat, R. M. and Reinhold, J. G., 1972. Zinc deficiency in man: The Shiraz experiment. *The American Journal of Medicine*, 53 (3), 277–284.

Han, Y., Gu, S., Oota, H., Osier, M. V., Pakstis, A. J., Speed, W. C., Kidd, J. R. and Kidd, K. K., 2007. Evidence of Positive Selection on a Class I ADH Locus. *The American Journal of Human Genetics*, 80 (3), 441–456.

Hancock, A. M., Alkorta-Aranburu, G., Witonsky, D. B. and Di Rienzo, A., 2010. Adaptations to new environments in humans: the role of subtle allele frequency shifts. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365 (1552), 2459–2468.

Hancock, A. M., Clark, V. J., Qian, Y. and Di Rienzo, A., 2011. Population Genetic Analysis of the Uncoupling Proteins Supports a Role for UCP3 in Human Cold Resistance. *Molecular Biology and Evolution*, 28 (1), 601–614.

Hancock, A. M., Witonsky, D. B., Alkorta-Aranburu, G., Beall, C. M., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J. K., Coop, G. and Rienzo, A. D., 2011. Adaptations to Climate-Mediated Selective Pressures in Humans. *PLOS Genetics*, 7 (4), e1001375.

Hancock, A. M., Witonsky, D. B., Ehler, E., Alkorta-Aranburu, G., Beall, C., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J., Coop, G. and Di Rienzo, A., 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences*, 107 (supplement_2), 8924–8930.

Hancock, A. M., Witonsky, D. B., Gordon, A. S., Eshel, G., Pritchard, J. K., Coop, G. and Rienzo, A. D., 2008. Adaptations to Climate in Candidate Genes for Common Metabolic Disorders. *PLOS Genetics*, 4 (2), e32.

Harris, K. and Nielsen, R., 2013. Inferring Demographic History from a Spectrum of Shared Haplotype Lengths. *PLOS Genetics*, 9 (6), e1003521.

Harris, K. and Pritchard, J. K., 2017. Rapid evolution of the human mutation spectrum. *eLife*, 6, e24284.

Hassani, A., Azapagic, A. and Shokri, N., 2021. Global predictions of primary soil salinization under changing climate in the 21st century. *Nature Communications*, 12 (1), 6663.

Hatfield, D. L., Carlson, B. A., Xu, X.-M., Mix, H. and Gladyshev, V. N., 2006. Selenocysteine incorporation machinery and the role of selenoproteins in development and health. *Progress in Nucleic Acid Research and Molecular Biology*, 81, 97–142.

Hawkes, C. F. C., 1949. The Dawn of European civilization. *Nature*, 163 (4151), 785–785.

He, X., Augusto, L., Goll, D. S., Ringeval, B., Wang, Y., Helfenstein, J., Huang, Y., Yu, K., Wang, Z., Yang, Y. and Hou, E., 2021. Global patterns and drivers of soil total phosphorus concentration. *Earth System Science Data*, 13 (12), 5831–5846.

He, Z., Xu, S. and Shi, S., 2020. Adaptive convergence at the genomic level—prevalent, uncommon or very rare? *National Science Review*, 7 (6), 947–951.

Heard, E. and Martienssen, R. A., 2014. Transgenerational Epigenetic Inheritance: myths and mechanisms. *Cell*, 157 (1), 95–109.

Heath, K. M., Axton, J. H., McCullough, J. M. and Harris, N., 2016. The evolutionary adaptation of the C282Y mutation to culture and climate during the European Neolithic. *American Journal of Physical Anthropology*, 160 (1), 86–101.

Hedges, S. B., 2002. The origin and evolution of model organisms. *Nature Reviews Genetics*, 3 (11), 838–849.

Hefnawy, A. E. G. and Tórtora-Pérez, J. L., 2010. The importance of selenium and the effects of its deficiency in animal health. *Small Ruminant Research*, 89 (2), 185–192.

Hejase, H. A., Dukler, N. and Siepel, A., 2020. From Summary Statistics to Gene Trees: Methods for Inferring Positive Selection. *Trends in Genetics*, 36 (4), 243–258.

Hengl, T., Leenaars, J. G. B., Shepherd, K. D., Walsh, M. G., Heuvelink, G. B. M., Mamo, T., Tilahun, H., Berkhout, E., Cooper, M., Fegraus, E., Wheeler, I. and Kwabena, N. A., 2017. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. *Nutrient Cycling in Agroecosystems*, 109 (1), 77–102.

Henn, B. M., Gignoux, C. R., Jobin, M., Granka, J. M., Macpherson, J. M., Kidd, J. M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A., Lin, A. A., Underhill, P. A., Comas, D., Kidd, K. K., Norman, P. J., Parham, P., Bustamante, C. D., Mountain, J. L. and Feldman, M.

W., 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences*, 108 (13), 5154–5162.
Herbette, S., Roeckel-Drevet, P. and Drevet, J. R., 2007. Seleno-independent glutathione peroxidases. *The FEBS Journal*, 274 (9), 2163–2180.
Hermisson, J. and Pennings, P. S., 2005. Soft Sweeps: Molecular Population Genetics of Adaptation From Standing Genetic Variation. *Genetics*, 169 (4), 2335–2352.
Hermisson, J. and Pennings, P. S., 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution*, 8 (6), 700–716.
Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., 1000 GENOMES PROJECT, Sella, G. and Przeworski, M., 2011. Classic Selective Sweeps Were Rare in Recent Human Evolution. *Science*, 331 (6019), 920–924.
Herráez, D. L., Bauchet, M., Tang, K., Theunert, C., Pugach, I., Li, J., Nandineni, M. R., Gross, A., Scholz, M. and Stoneking, M., 2009. Genetic Variation and Recent Positive Selection in Worldwide Human Populations: Evidence from Nearly 1 Million SNPs. *PLOS ONE*, 4 (11), e7888.
Hershkovitz, I., Marder, O., Ayalon, A., Bar-Matthews, M., Yasur, G., Boaretto, E., Caracuta, V., Alex, B., Frumkin, A., . . . Barzilai, O., 2015. Levantine cranium from Manot Cave (Israel) foreshadows the first European modern humans. *Nature*, 520 (7546), 216–219.
Hesse, F. G., 1959. A Dietary Study of the Pima Indian. *The American Journal of Clinical Nutrition*, 7 (5), 532–537.
Higdon, J. W., Bininda-Emonds, O. R., Beck, R. M. and Ferguson, S. H., 2007. Phylogeny and divergence of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. *BMC Evolutionary Biology*, 7 (1), 216.
Higham, T., Douka, K., Wood, R., Ramsey, C. B., Brock, F., Basell, L., Camps, M., Arrizabalaga, A., Baena, J., . . . Jacobi, R., 2014. The timing and spatiotemporal patterning of Neanderthal disappearance. *Nature*, 512 (7514), 306–309.
Hill, A. V. S., Allsopp, C. E. M., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J. and Greenwood, B. M., 1991. Common West African HLA antigens are associated with protection from severe malaria. *Nature*, 352 (6336), 595–600.
Hirbo, J. B., Ranciaro, A. and Tishkoff, S. A., 2012. Population structure and migration in Africa: correlations between archaeological, linguistic, and genetic data. *In*: Campbell, B. C. and Crawford, M. H., eds. *Causes and Consequences of Human Migration: An Evolutionary Perspective* [online]. Cambridge: Cambridge University Press, 135–171. Available from: https://www.cambridge.org/core/books/causes-and-consequences-of-human-migration/population-structure-and-migration-in-africa-correlations-between-archaeological-linguistic-and-genetic-data/B463E94F96A23B8DD4EB0C60AC99A383 [Accessed 14 Mar 2023].
Horning, K. J., Caito, S. W., Tipps, K. G., Bowman, A. B. and Aschner, M., 2015. Manganese Is Essential for Neuronal Health. *Annual Review of Nutrition*, 35, 71–108.
Hornung, T. C. and Biesalski, H.-K., 2019. Glut-1 explains the evolutionary advantage of the loss of endogenous vitamin C-synthesis: The electron transfer hypothesis. *Evolution, Medicine, and Public Health*, 2019 (1), 221–231.
Hou, X., Zhang, X., Li, X., Huang, T., Li, W., Zhang, H., Huang, H. and Wen, Y., 2022. Genomic insights into the genetic structure and population history of Mongolians in Liaoning Province. *Frontiers in Genetics* [online], 13. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2022.947758 [Accessed 30 Jan 2023].

Houillier, P., 2014. Mechanisms and regulation of renal magnesium transport. *Annual Review of Physiology*, 76, 411–430.

Huang, Q. Q., Sallah, N., Dunca, D., Trivedi, B., Hunt, K. A., Hodgson, S., Lambert, S. A., Arciero, E., Wright, J., . . . Kuchenbaecker, K., 2022. Transferability of genetic loci and polygenic scores for cardiometabolic traits in British Pakistani and Bangladeshi individuals. *Nature Communications*, 13 (1), 4664.

Hubisz, M. J., Williams, A. L. and Siepel, A., 2020. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. *PLoS genetics*, 16 (8), e1008895.

Hubisz, M. and Siepel, A., 2020. Inference of Ancestral Recombination Graphs Using ARGweaver. *In*: Dutheil, J. Y., ed. *Statistical Population Genomics* [online]. New York, NY: Springer US, 231–266. Available from: https://doi.org/10.1007/978-1-0716-0199-0_10 [Accessed 25 Jan 2023].

Huchon, D., Madsen, O., Sibbald, M. J. J. B., Ament, K., Stanhope, M. J., Catzeflis, F., de Jong, W. W. and Douzery, E. J. P., 2002. Rodent Phylogeny and a Timescale for the Evolution of Glires: Evidence from an Extensive Taxon Sampling Using Three Nuclear Genes. *Molecular Biology and Evolution*, 19 (7), 1053–1065.

Huerta-Sánchez, E., Jin, X., Asan, Bianba, Z., Peter, B. M., Vinckenbosch, N., Liang, Y., Yi, X., He, M., . . . Nielsen, R., 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature*, 512 (7513), 194–197.

Hughes, A. L., 1997. The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 256 (1346), 119–124.

Hughes, D. A., Tang, K., Strotmann, R., Schöneberg, T., Prenen, J., Nilius, B. and Stoneking, M., 2008. Parallel Selection on TRPV6 in Human Populations. *PLOS ONE*, 3 (2), e1686.

Hunter, R. W., Dhaun, N. and Bailey, M. A., 2022. The impact of excessive salt intake on human health. *Nature Reviews Nephrology*, 18 (5), 321–335.

Hurley, S. W. and Johnson, A. K., 2015. The biopsychology of salt hunger and sodium deficiency. *Pflügers Archiv - European Journal of Physiology*, 467 (3), 445–456.

Hurst, R., Siyame, E. W. P., Young, S. D., Chilimba, A. D. C., Joy, E. J. M., Black, C. R., Ander, E. L., Watts, M. J., Chilima, B., Gondwe, J., Kang'ombe, D., Stein, A. J., Fairweather-Tait, S. J., Gibson, R. S., Kalimbira, A. A. and Broadley, M. R., 2013. Soil-type influences human selenium status and underlies widespread selenium deficiency risks in Malawi. *Scientific Reports*, 3 (1), 1425.

Huxley, J., 1942. *Evolution: The Modern Synthesis*. London: Allen & Unwin.

Hwang, J.-Y., Lee, S. H., Go, M. J., Kim, B.-J., Kou, I., Ikegawa, S., Guo, Y., Deng, H.-W., Raychaudhuri, S., Kim, Y. J., . . . Koh, J.-M., 2013. Meta-analysis identifies a MECOM gene as a novel predisposing factor of osteoporotic fracture. *Journal of Medical Genetics*, 50 (4), 212–219.

Ibrahim, S. A. Z., Kerkadi, A. and Agouni, A., 2019. Selenium and Health: An Update on the Situation in the Middle East and North Africa. *Nutrients*, 11 (7), 1457.

Ilardo, M. A., Moltke, I., Korneliussen, T. S., Cheng, J., Stern, A. J., Racimo, F., Damgaard, P. de B., Sikora, M., Seguin-Orlando, . . . Willerslev, E., 2018. Physiological and Genetic Adaptations to Diving in Sea Nomads. *Cell*, 173 (3), 569-580.e15.

Ilardo, M. and Nielsen, R., 2018. Human adaptation to extreme environmental conditions. *Current Opinion in Genetics & Development*, 53, 77–82.

Ishfaq, M., Wakeel, A., Shahzad, M. N., Kiran, A. and Li, X., 2021. Severity of zinc and iron malnutrition linked to low intake through a staple crop: a case study in east-central Pakistan. *Environmental Geochemistry and Health*, 43 (10), 4219–4233.

Issaka, R. N., Masunaga, T., Kosaki, T. and Wakatsuki, T., 1996. Soils of inland valleys of West Africa: General fertility parameters. *Soil Science and Plant Nutrition*, 42 (1), 71–80.

Jacobs, G. S., Hudjashov, G., Saag, L., Kusuma, P., Darusallam, C. C., Lawson, D. J., Mondal, M., Pagani, L., Ricaut, F.-X., Stoneking, M., Metspalu, M., Sudoyo, H., Lansing, J. S. and Cox, M. P., 2019. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*, 177 (4), 1010-1021.e32.

Jacobs, L. C., Wollstein, A., Lao, O., Hofman, A., Klaver, C. C., Uitterlinden, A. G., Nijsten, T., Kayser, M. and Liu, F., 2013. Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. *Human Genetics*, 132 (2), 147–158.

Jain, G., Ong, S. and Warnock, D. G., 2013. Genetic Disorders of Potassium Homeostasis. *Seminars in Nephrology*, 33 (3), 300–309.

Jakupoglu, C., Przemeck, G. K. H., Schneider, M., Moreno, S. G., Mayr, N., Hatzopoulos, A. K., de Angelis, M. H., Wurst, W., Bornkamm, G. W., Brielmeier, M. and Conrad, M., 2005. Cytoplasmic thioredoxin reductase is essential for embryogenesis but dispensable for cardiac development. *Molecular and Cellular Biology*, 25 (5), 1980–1988.

Jayaraman, V., Toledo-Patiño, S., Noda-García, L. and Laurino, P., 2022. Mechanisms of protein evolution. *Protein Science*, 31 (7), e4362.

Jensen, R. A., 1976. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology*, 30 (1), 409–425.

Johansson, L., Gafvelin, G. and Arnér, E. S. J., 2005. Selenocysteine in proteins—properties and biotechnological use. *Biochimica et Biophysica Acta (BBA) - General Subjects*, 1726 (1), 1–13.

Jones, G. D., Droz, B., Greve, P., Gottschalk, P., Poffet, D., McGrath, S. P., Seneviratne, S. I., Smith, P. and Winkel, L. H. E., 2017. Selenium deficiency risk predicted to increase under future climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 114 (11), 2848–2853.

Jones, D. T., Taylor, W. R., Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences, *Bioinformatics*, 8 (3).

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., . . . Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596 (7873), 583–589.

Juric, I., Aeschbacher, S. and Coop, G., 2016. The Strength of Selection against Neanderthal Introgression. *PLOS Genetics*, 12 (11), e1006340.

Kambe, T., Tsuji, T., Hashimoto, A. and Itsumura, N., 2015. The Physiological, Biochemical, and Molecular Roles of Zinc Transporters in Zinc Homeostasis and Metabolism. *Physiological Reviews*, 95 (3), 749–784.

Kamberov, Y. G., Wang, S., Tan, J., Gerbault, P., Wark, A., Tan, L., Yang, Y., Li, S., Tang, K., . . .Sabeti, P. C., 2013. Modeling Recent Human Evolution in Mice by Expression of a Selected EDAR Variant. *Cell*, 152 (4), 691–702.

Kanzok, S. M., Fechner, A., Bauer, H., Ulschmid, J. K., Müller, H.-M., Botella-Munoz, J., Schneuwly, S., Schirmer, R. H. and Becker, K., 2001. Substitution of the Thioredoxin System for Glutathione Reductase in Drosophila melanogaster. *Science*, 291 (5504), 643–646.

Kaplan, N. L., Hudson, R. R. and Langley, C. H., 1989. The 'hitchhiking effect' revisited. *Genetics*, 123 (4), 887–899.

Karim, Md. R., Zhang, Y.-Q., Tian, D., Chen, F.-J., Zhang, F.-S. and Zou, C.-Q., 2012. Genotypic differences in zinc efficiency of Chinese maize evaluated in a pot experiment. *Journal of the Science of Food and Agriculture*, 92 (12), 2552–2559.

Karimov, A., Qadir, M., Noble, A., Vyshpolsky, F. and Anzelm, K., 2009. Development of Magnesium-Dominant Soils Under Irrigated Agriculture in Southern Kazakhstan. *Pedosphere*, 19 (3), 331–343.

Karlsson, E. K., Kwiatkowski, D. P. and Sabeti, P. C., 2014. Natural selection and infectious disease in human populations. *Nature Reviews Genetics*, 15 (6), 379–393.

Kataoka, K., Fujita, H., Isa, M., Gotoh, S., Arasaki, A., Ishida, H., Kimura, R. 2021.The human *EDAR* 370V/A polymorphism affects tooth root morphology potentially through the modification of a reaction–diffusion system. *Sci Rep* **11**, 5143

Katoh, K., Rozewicki, J. and Yamada, K. D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, 20 (4), 1160–1166.

Kaur, H. and Garg, N., 2021. Zinc toxicity in plants: a review. *Planta*, 253 (6), 129.

Keats, E. C., Neufeld, L. M., Garrett, G. S., Mbuya, M. N. N. and Bhutta, Z. A., 2019. Improved micronutrient status and health outcomes in low- and middle-income countries following large-scale fortification: evidence from a systematic review and meta-analysis. *The American Journal of Clinical Nutrition*, 109 (6), 1696–1708.

Keefer, R. F., 1999. Micronutrients. *In*: Keefer, R. F., ed. *Handbook of Soils for Landscape Architects* [online]. Oxford University Press, 0. Available from: https://doi.org/10.1093/oso/9780195121025.003.0016 [Accessed 9 Mar 2023].

Keightley, P. D. and Eyre-Walker, A., 2010. What can we learn about the distribution of fitness effects of new mutations from DNA sequence data? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365 (1544), 1187–1193.

Keinan, A., Mullikin, J. C., Patterson, N. and Reich, D., 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nature Genetics*, 39 (10), 1251–1255.

Kelleher, J., Etheridge, A. M. and McVean, G., 2016. Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes. *PLOS Computational Biology*, 12 (5), e1004842.

Kelleher, J., Wong, Y., Wohns, A. W., Fadil, C., Albers, P. K. and McVean, G., 2019. Inferring whole-genome histories in large population datasets. *Nature Genetics*, 51 (9), 1330–1338.

Kelly, F. C. and Snedden, W. W., 1960. Prevalence and geographical distribution of endemic goitre. *Monograph Series. World Health Organization*, 44, 27–233.

Kern, A. D., Schrider, D. R. 2018. diploS/HIC: An Updated Approach to Classifying Selective Sweeps, *G3 Genes|Genomes|Genetics* 8(6).

Kestenbaum, B., Glazer, N. L., Köttgen, A., Felix, J. F., Hwang, S.-J., Liu, Y., Lohman, K., Kritchevsky, S. B., Hausman, D. B., . . . Fox, C. S., 2010. Common Genetic Variants Associate with Serum Phosphorus Concentration. *Journal of the American Society of Nephrology : JASN*, 21 (7), 1223–1232.

Key, F. M., Abdul-Aziz, M. A., Mundry, R., Peter, B. M., Sekar, A., D'Amato, M., Dennis, M. Y., Schmidt, J. M. and Andrés, A. M., 2018. Human local adaptation of the TRPM8 cold receptor along a latitudinal cline. *PLOS Genetics*, 14 (5), e1007298.

Key, F. M., Fu, Q., Romagne, F., Lachmann, M. and Andres, A. M., 2016. Human adaptation and population differentiation in the light of ancient genomes. *Nature Communications*, 7.

Key, F. M., Peter, B., Dennis, M. Y., Huerta-Sánchez, E., Tang, W., Prokunina-Olsson, L., Nielsen, R. and Andrés, A. M., 2014. Selection on a Variant Associated with Improved Viral Clearance Drives Local, Adaptive Pseudogenization of Interferon Lambda 4 (IFNL4). *PLOS Genetics*, 10 (10), e1004681.

Khan, S. T., Malik, A., Alwarthan, A. and Shaik, M. R., 2022. The enormity of the zinc deficiency problem and available solutions; an overview. *Arabian Journal of Chemistry*, 15 (3), 103668.

Khanal, R. C. and Nemere, I., 2008. Endocrine regulation of calcium transport in epithelia. *Clinical and Experimental Pharmacology & Physiology*, 35 (11), 1277–1287.

Khoshgoftarmanesh, A. H., Schulin, R., Chaney, R. L., Daneshbakhsh, B. and Afyuni, M., 2010. Micronutrient-efficient genotypes for crop yield and nutritional quality in sustainable agriculture. A review. *Agronomy for Sustainable Development*, 30 (1), 83–107.

Kihara, J., Bolo, P., Kinyua, M., Rurinda, J. and Piikki, K., 2020. Micronutrient deficiencies in African soils and the human nutritional nexus: opportunities with staple crops. *Environmental Geochemistry and Health*, 42 (9), 3015–3033.

Kilpinen, H., Goncalves, A., Leha, A., Afzal, V., Alasoo, K., Ashford, S., Bala, S., Bensaddek, D., Casale, F. P., . . . Gaffney, D. J., 2017. Common genetic variation drives molecular heterogeneity in human iPSCs. *Nature*, 546 (7658), 370–375.

Kim, B. Y., Huber, C. D. and Lohmueller, K. E., 2017. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics*, 206 (1), 345–361.

Kim, B. Y., Huber, C. D. and Lohmueller, K. E., 2018. Deleterious variation shapes the genomic landscape of introgression. *PLOS Genetics*, 14 (10), e1007741.

Kim, M.-J., Lee, B. C., Hwang, K. Y., Gladyshev, V. N. and Kim, H.-Y., 2015. Selenium utilization in thioredoxin and catalytic advantage provided by selenocysteine. *Biochemical and biophysical research communications*, 461 (4), 648–652.

Kimura, M., 1968. Evolutionary Rate at the Molecular Level. *Nature*, 217 (5129), 624–626.

Kimura, M. and Ohta, T., 1969. The Average Number of Generations until Fixation of a Mutant Gene in a Finite Population. *Genetics*, 61 (3), 763–771.

Köhrle, J., 2000. The deiodinase family: selenoenzymes regulating thyroid hormone availability and action. *Cellular and molecular life sciences: CMLS*, 57 (13–14), 1853–1863.

Koivistoinen, P. and Huttunen, J. K., 1986. Selenium in food and nutrition in Finland. An overview on research and action. *Annals of Clinical Research*, 18 (1), 13–17.

Korfmann, K., Gaggiotti, G., Fumagalli, M. 2023. Deep Learning in Population Genetics, *Genome Biology and Evolution,* 15 (2).

Kovacs, G., Montalbetti, N., Franz, M.-C., Graeter, S., Simonin, A. and Hediger, M. A., 2013. Human TRPV5 and TRPV6: key players in cadmium and zinc toxicity. *Cell Calcium*, 54 (4), 276–286.

Kryukov, G. V., Castellano, S., Novoselov, S. V., Lobanov, A. V., Zehtab, O., Guigó, R. and Gladyshev, V. N., 2003. Characterization of Mammalian Selenoproteomes. *Science*, 300 (5624), 1439–1443.

Kuhlwilm, M., Gronau, I., Hubisz, M. J., de Filippo, C., Prado-Martinez, J., Kircher, M., Fu, Q., Burbano, H. A., Lalueza-Fox, C., . . . Castellano, S., 2016. Ancient gene flow from early modern humans into Eastern Neanderthals. *Nature*, 530 (7591), 429–433.

Kwiatkowski, D. P., 2005. How Malaria Has Affected the Human Genome and What Human Genetics Can Teach Us about Malaria. *The American Journal of Human Genetics*, 77 (2), 171–192.

Labunskyy, V. M., Hatfield, D. L. and Gladyshev, V. N., 2014. Selenoproteins: Molecular Pathways and Physiological Roles. *Physiological Reviews*, 94 (3), 739–777.

Laekemariam, F., Kibret, K. and Shiferaw, H., 2018. Potassium (K)-to-magnesium (Mg) ratio, its spatial variability and implications to potential Mg-induced K deficiency in Nitisols of Southern Ethiopia. *Agriculture & Food Security*, 7 (1), 13.

Lahr, M. M. and Foley, R., 1994. Multiple dispersals and modern human origins. *Evolutionary Anthropology: Issues, News, and Reviews*, 3 (2), 48–60.

Laland, K. N., Uller, T., Feldman, M. W., Sterelny, K., Müller, G. B., Moczek, A., Jablonka, E. and Odling-Smee, J., 2015. The extended evolutionary synthesis: its structure, assumptions and predictions. *Proceedings. Biological Sciences*, 282 (1813), 20151019.

Lamason, R. L., Mohideen, M.-A. P. K., Mest, J. R., Wong, A. C., Norton, H. L., Aros, M. C., Jurynec, M. J., Mao, X., Humphreville, V. R., . . . Cheng, K. C., 2005. SLC24A5, a putative cation exchanger, affects pigmentation in zebrafish and humans. *Science (New York, N.Y.)*, 310 (5755), 1782–1786.

Larsson, D. G. J. and Flach, C.-F., 2022. Antibiotic resistance in the environment. *Nature Reviews Microbiology*, 20 (5), 257–269.

Latham, K. J., 2013. Human Health and the Neolithic Revolution: an Overview of Impacts of the Agricultural Transition on Oral Health, Epidemiology, and the Human Body. *Nebraska Anthropologist*, 187.

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., Sudmant, P. H., Schraiber, J. G., Castellano, S . . . Krause, J., 2014. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513 (7518), 409–413.

Le Corre, V. and Kremer, A., 2012. The genetic differentiation at quantitative trait loci under local adaptation. *Molecular Ecology*, 21 (7), 1548–1566.

LeCun, Y., Bengio, Y., Hinton, G. Deep learning. 2015. *Nature* 521, 436–444

Le, M. K., Smith, O. S., Akbari, A., Harpak, A., Reich, D. and Narasimhan, V. M., 2022. *1,000 ancient genomes uncover 10,000 years of natural selection in Europe* [online]. Genomics. preprint. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.08.24.505188 [Accessed 15 Mar 2023].

Lee, S. R., Bar-Noy, S., Kwon, J., Levine, R. L., Stadtman, T. C. and Rhee, S. G., 2000. Mammalian thioredoxin reductase: oxidation of the C-terminal cysteine/selenocysteine active site forms a thioselenide, and replacement of selenium with sulfur markedly reduces catalytic activity. *Proceedings of the National Academy of Sciences of the United States of America*, 97 (6), 2521–2526.

Leigh, J. W. and Bryant, D., 2015. popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, 6 (9), 1110–1116.

Lenski, R. E., Ofria, C., Pennock, R. T. and Adami, C., 2003. The evolutionary origin of complex features. *Nature*, 423 (6936), 139–144.

Levi, S., Ripamonti, M., Dardi, M., Cozzi, A. and Santambrogio, P., 2021. Mitochondrial Ferritin: Its Role in Physiological and Pathological Conditions. *Cells*, 10 (8), 1969.

Lewis, J. J., Van Belleghem, S. M., Papa, R., Danko, C. G. and Reed, R. D., 2020. Many functionally connected loci foster adaptive diversification along a neotropical hybrid zone. *Science Advances*, 6 (39), eabb8617.

Li, H., Mukherjee, N., Soundararajan, U., Tarnok, Z., Barta, C., Khaliq, S., Mohyuddin, A., Kajuna, S.L., Mehdi, S.Q., Kidd, J.R., Kidd, K.K. 2007. Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. *Am J Hum Genet*. 81(4):842-6.

Li, D., Li, Y., Li, M., Che, T., Tian, S., Chen, B., Zhou, X., Zhang, G., Gaur, U., . . . Li, M., 2019. Population genomics identifies patterns of genetic diversity and selection in chicken. *BMC Genomics*, 20 (1), 263.

Li, D. and Zhang, J., 2014. Diet Shapes the Evolution of the Vertebrate Bitter Taste Receptor Gene Repertoire. *Molecular Biology and Evolution*, 31 (2), 303–309.

Liang, M. and Nielsen, R., 2014. The Lengths of Admixture Tracts. *Genetics*, 197 (3), 953–967.

Librado, P. and Orlando, L., 2018. Detecting Signatures of Positive Selection along Defined Branches of a Population Tree Using LSD. *Molecular Biology and Evolution*, 35 (6), 1520–1535.

Ligowe, I. S., Phiri, F. P., Ander, E. L., Bailey, E. H., Chilimba, A. D. C., Gashu, D., Joy, E. J. M., Lark, R. M., Kabambe, V., . . . Broadley, M. R., 2020. Selenium deficiency risks in sub-Saharan African food systems and their geospatial linkages. *Proceedings of the Nutrition Society*, 79 (4), 457–467.

Lim, E. T., Würtz, P., Havulinna, A. S., Palta, P., Tukiainen, T., Rehnström, K., Esko, T., Mägi, R., Inouye, M., . . . Project, for the S. I. S. (SISu), 2014. Distribution and Medical Impact of Loss-of-Function Variants in the Finnish Founder Population. *PLOS Genetics*, 10 (7), e1004494.

Lindo, J., Huerta-Sánchez, E., Nakagome, S., Rasmussen, M., Petzelt, B., Mitchell, J., Cybulski, J. S., Willerslev, E., DeGiorgio, M. and Malhi, R. S., 2016. A time transect of exomes from a Native American population before and after European contact. *Nature Communications*, 7 (1), 13175.

Lindsay, G. B. and Edwards, G., 1988. Creating effective health coalitions. *Health education*, 19 (4), 35–36.

Liu, H., Yin, L., Board, P. G., Han, X., Fan, Z., Fang, J., Lu, Z., Zhang, Y. and Wei, J., 2012. Expression of selenocysteine-containing glutathione S-transferase in eukaryote. *Protein Expression and Purification*, 84 (1), 59–63.

Liu, W., Martinón-Torres, M., Cai, Y., Xing, S., Tong, H., Pei, S., Sier, M. J., Wu, X., Edwards, R. L., Cheng, H., Li, Y., Yang, X., de Castro, J. M. B. and Wu, X., 2015. The earliest unequivocally modern humans in southern China. *Nature*, 526 (7575), 696–699.

Liu, X., Fu, Y.-X., Maxwell, T. J. and Boerwinkle, E., 2010. Estimating population genetic parameters and comparing model goodness-of-fit using DNA sequences with error. *Genome Research*, 20 (1), 101–109.

Liu, Y., Tian, X., Liu, R., Liu, S. and Zuza, A. V., 2021. Key driving factors of selenium-enriched soil in the low-Se geological belt: A case study in Red Beds of Sichuan Basin, China. *CATENA*, 196, 104926.

Lobanov, A. V., Fomenko, D. E., Zhang, Y., Sengupta, A., Hatfield, D. L. and Gladyshev, V. N., 2007. Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biology*, 8 (9), R198.

Loeb, D. D., Swanstrom, R., Everitt, L., Manchester, M., Stamper, S. E. and Hutchison, C. A., 1989. Complete mutagenesis of the HIV-1 protease. *Nature*, 340 (6232), 397–400.

Loewe, L. and Hill, W. G., 2010. The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365 (1544), 1153–1167.

López, S., van Dorp, L. and Hellenthal, G., 2016. Human Dispersal Out of Africa: A Lasting Debate. *Evolutionary Bioinformatics Online*, 11 (Suppl 2), 57–68.

Luis, J. R., Rowold, D. J., Regueiro, M., Caeiro, B., Cinnioğlu, C., Roseman, C., Underhill, P. A., Cavalli-Sforza, L. L. and Herrera, R. J., 2004. The Levant versus the Horn of Africa: Evidence for Bidirectional Corridors of Human Migrations. *American Journal of Human Genetics*, 74 (3), 532–544.

Lunzer, M., Golding, G. B. and Dean, A. M., 2010. Pervasive Cryptic Epistasis in Molecular Evolution. *PLOS Genetics*, 6 (10), e1001162.

Lynch, M. and Ho, W.-C., 2020. The Limits to Estimating Population-Genetic Parameters with Temporal Data. *Genome Biology and Evolution*, 12 (4), 443–455.

Lyons, G., 2018. Biofortification of Cereals With Foliar Selenium and Iodine Could Reduce Hypothyroidism. *Frontiers in Plant Science*, 9, 730.

Ma, G., Jin, Y., Li, Y., Zhai, F., Kok, F. J., Jacobsen, E. and Yang, X., 2008. Iron and zinc deficiencies in China: what is a feasible and cost-effective strategy? *Public Health Nutrition*, 11 (6), 632–638.

Ma, Y., Ding, X., Qanbari, S., Weigend, S., Zhang, Q. and Simianer, H., 2015. Properties of different selection signature statistics and a new strategy for combining them. *Heredity*, 115 (5), 426–436.

Maca-Meyer, N., González, A. M., Larruga, J. M., Flores, C. and Cabrera, V. M., 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC genetics*, 2, 13.

MacFarquhar, J. K., Broussard, D. L., Melstrom, P., Hutchinson, R., Wolkin, A., Martin, C., Burk, R. F., Dunn, J. R., Green, A. L., Hammond, R., Schaffner, W. and Jones, T. F., 2010. Acute Selenium Toxicity Associated With a Dietary Supplement. *Archives of Internal Medicine*, 170 (3), 256–261.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D. and Ravikesavan, R., 2013. Gene duplication as a major force in evolution. *Journal of Genetics*, 92 (1), 155–161.

Maisnier-Patin, S. and Andersson, D. I., 2004. Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. *Research in Microbiology*, 155 (5), 360–369.

Manning, L., Laman, M., Rosanas-Urgell, A., Michon, P., Aipit, S., Bona, C., Siba, P., Mueller, I. and Davis, T. M. E., 2012. Severe Anemia in Papua New Guinean Children from a Malaria-Endemic Area: A Case-Control Etiologic Study. *PLoS Neglected Tropical Diseases*, 6 (12), e1972.

Manus, M. B., 2018. Evolutionary mismatch. *Evolution, Medicine, and Public Health*, 2018 (1), 190–191.

Marciniak, S. and Perry, G. H., 2017. Harnessing ancient genomes to study the history of human adaptation. *Nature Reviews Genetics*, 18 (11), 659–674.

Mariotti, M., Ridge, P. G., Zhang, Y., Lobanov, A. V., Pringle, T. H., Guigo, R., Hatfield, D. L. and Gladyshev, V. N., 2012. Composition and Evolution of the Vertebrate and Mammalian Selenoproteomes. *PLOS ONE*, 7 (3), e33066.

Markadieu, N. and Delpire, E., 2014. Physiology and Pathophysiology of SLC12A1/2 transporters. *Pflugers Archiv : European journal of physiology*, 466 (1), 10.1007/s00424-013-1370–5.

Mather, K., Moran, P. A. P. and Smith, C. A. B., 1967. Commentary on R. A. Fisher's paper on The Correlation Between Relatives on the Supposition of Mendelian Inheritance. *Population Studies*, 20 (3), 372.

Mathieson, I., 2021. The omnigenic model and polygenic prediction of complex traits. *The American Journal of Human Genetics*, 108 (9), 1558–1563.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., . . . Reich, D., 2018. The genomic history of southeastern Europe. *Nature*, 555 (7695), 197–203.

Mathieson, I., Lazaridis, I., Rohland, N., Mallick, S., Patterson, N., Roodenberg, S. A., Harney, E., Stewardson, K., Fernandes, D., . . . Reich, D., 2015. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528 (7583), 499–503.

Mathieson, I. and Terhorst, J., 2022. Direct detection of natural selection in Bronze Age Britain. *Genome Res.* doi:10.1101/gr.276862.122

Mathieson, S. and Mathieson, I., 2018. FADS1 and the Timing of Human Adaptation to Agriculture. *Molecular Biology and Evolution*, 35 (12), 2957–2970.

Matsui, M., Oshima, M., Oshima, H., Takaku, K., Maruyama, T., Yodoi, J. and Taketo, M. M., 1996. Early embryonic lethality caused by targeted disruption of the mouse thioredoxin gene. *Developmental Biology*, 178 (1), 179–185.

Mauro, A. A. and Ghalambor, C. K., 2020. Trade-offs, Pleiotropy, and Shared Molecular Pathways: A Unified View of Constraints on Adaptation. *Integrative and Comparative Biology*, 60 (2), 332–347.

May, T. W., Fairchild, J. F., Petty, J. D., Walther, M. J., Lucero, J., Delvaux, M., Manring, J. and Armbruster, M., 2008. An evaluation of selenium concentrations in water, sediment, invertebrates, and fish from the Solomon River Basin. *Environmental Monitoring and Assessment*, 137 (1–3), 213–232.

McCarthy, R. C. and Lucas, L., 2014. A morphometric re-assessment of BOU-VP-16/1 from Herto, Ethiopia. *Journal of Human Evolution*, 74, 114–117.

McDougall, I., Brown, F. H. and Fleagle, J. G., 2005. Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature*, 433 (7027), 733–736.

Mcgeorge, W. T., n.d. FACTORS INFLUENCING THE AVAILABILITY OF NATIVE SOIL PHOSPHATE AND PHOSPHATE FERTILIZERS IN ARIZONA SOILS.

McManus, K. F., Taravella, A. M., Henn, B. M., Bustamante, C. D., Sikora, M. and Cornejo, O. E., 2017. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. *PLOS Genetics*, 13 (3), e1006560.

McWilliams, S. R., 2011. Ecology of Vertebrate Nutrition. *In*: *eLS* [online]. John Wiley & Sons, Ltd. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0003211.pub2 [Accessed 13 Mar 2023].

Mehdi, Y. and Dufrasne, I., 2016. Selenium in Cattle: A Review. *Molecules*, 21 (4), 545.

Mehdi, Y., Hornick, J.-L., Istasse, L. and Dufrasne, I., 2013. Selenium in the Environment, Metabolism and Involvement in Body Functions. *Molecules*, 18 (3), 3292–3311.

Mertz, W., 1981. The Essential Trace Elements. *Science*, 213 (4514), 1332–1338.

Meyer, M., Kircher, M., Gansauge, M.-T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., . . . Pääbo, S., 2012. A high-coverage genome sequence from an archaic Denisovan individual. *Science (New York, N.Y.)*, 338 (6104), 222–226.

Miles, M., 1998. Goitre, cretinism and iodine in South Asia: historical perspectives on a continuing scourge. *Medical History*, 42 (1), 47–67.

Minster, R. L., Hawley, N. L., Su, C.-T., Sun, G., Kershaw, E. E., Cheng, H., Buhule, O. D., Lin, J., Reupena, M. S., Viali, S., Tuitele, J., Naseri, T., Urban, Z., Deka, R., Weeks, D. E. and McGarvey, S. T., 2016. A thrifty variant in CREBRF strongly influences body mass index in Samoans. *Nature Genetics*, 48 (9), 1049–1054.

Mirmiran, P., Golzarand, M., Serra-Majem, L. and Azizi, F., 2012. Iron, Iodine and Vitamin A in the Middle East; A Systematic Review of Deficiency and Food Fortification. *Iranian Journal of Public Health*, 41 (8), 8–19.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D. and Bateman, A., 2021. Pfam: The protein families database in 2021. *Nucleic Acids Research*, 49 (D1), D412–D419.

Molgedey, L. and Schuster, H. G., 1994. Separation of a mixture of independent signals using time delayed correlations. *Physical Review Letters*, 72 (23), 3634–3637.

Mondal, M., Bertranpetit, J. and Lao, O., 2019. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. *Nature Communications*, 10 (1), 246.

Monteiro, J. P., Kussmann, M. and Kaput, J., 2015. The genomics of micronutrient requirements. *Genes & Nutrition*, 10 (4), 19.

Moran, C. and Chatterjee, K., 2015. Resistance to thyroid hormone due to defective thyroid receptor alpha. *Best Practice & Research Clinical Endocrinology & Metabolism*, 29 (4), 647–657.

Moreno-Mayar, J. V., Potter, B. A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J. A., Albrechtsen, A., . . . Willerslev, E., 2018. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, 553 (7687), 203–207.

Moreno-Reyes, R., Egrise, D., Nève, J., Pasteels, J. L. and Schoutens, A., 2001. Selenium deficiency-induced growth retardation is associated with an impaired bone metabolism and osteopenia. *Journal of Bone and Mineral Research: The Official Journal of the American Society for Bone and Mineral Research*, 16 (8), 1556–1563.

Moshe, A. and Pupko, T., 2019. Ancestral sequence reconstruction: accounting for structural information by averaging over replacement matrices. *Bioinformatics (Oxford, England)*, 35 (15), 2562–2568.

Muckenthaler, M. U., Galy, B. and Hentze, M. W., 2008. Systemic iron homeostasis and the iron-responsive element/iron-regulatory protein (IRE/IRP) regulatory network. *Annual Review of Nutrition*, 28, 197–213.

Mughal, M. R., DeGiorgio, M. 2019. Localizing and Classifying Adaptive Targets with Trend Filtered Regression, *Molecular Biology and Evolution*, 36(2), 252-270.

Naidu, L. G. K., Sidhu, S., Sarkar, D. and Ramamurthy, V., 2011. Emerging deficiency of potassium in soils and crops of India. *Karnataka J. Agric. Sci.*, 24.

National Academies of Sciences, E., Division, H. and M., Board, F. and N., Potassium, C. to R. the D. R. I. for S. and, Oria, M., Harrison, M. and Stallings, V. A., 2019. *Potassium: Dietary Reference Intakes for Toxicity* [online]. Dietary Reference Intakes for Sodium and Potassium. National Academies Press (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK545424/ [Accessed 19 Mar 2023].

Naugler, C., 2008. Hemochromatosis: A Neolithic adaptation to cereal grain diets. *Medical Hypotheses*, 70 (3), 691–692.

Nédélec, Y., Sanz, J., Baharian, G., Szpiech, Z. A., Pacis, A., Dumaine, A., Grenier, J.-C., Freiman, A., Sams, A. J., . . . Barreiro, L. B., 2016. Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Pathogens. *Cell*, 167 (3), 657-669.e21.

Nell, J. P. and van Huyssteen, C. W., 2018. Prediction of primary salinity, sodicity and alkalinity in South African soils. *South African Journal of Plant and Soil*, 35 (3), 173–178.

Newcomb, R. D., Campbell, P. M., Ollis, D. L., Cheah, E., Russell, R. J. and Oakeshott, J. G., 1997. A single amino acid substitution converts a carboxylesterase to an organophosphorus hydrolase and confers insecticide resistance on a blowfly. *Proceedings of the National Academy of Sciences*, 94 (14), 7464–7468.

Nguyen, V. D., Saaranen, M. J., Karala, A.-R., Lappi, A.-K., Wang, L., Raykhel, I. B., Alanen, H. I., Salo, K. E. H., Wang, C. and Ruddock, L. W., 2011. Two Endoplasmic Reticulum PDI Peroxidases Increase the Efficiency of the Use of Peroxide during Disulfide Bond Formation. *Journal of Molecular Biology*, 406 (3), 503–515.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C. and Clark, A. G., 2007. Recent and ongoing selection in the human genome. *Nature reviews. Genetics*, 8 (11), 857–868.

Niepomniszcze, H., Bernatené, D. and Sartorio, G., 2009. Chapter 123 - Iodine Status in Individuals: An Argentine Perspective. *In*: Preedy, V. R., Burrow, G. N., and Watson, R., eds. *Comprehensive Handbook of Iodine* [online]. San Diego: Academic Press, 1191–1201. Available from:

https://www.sciencedirect.com/science/article/pii/B9780123741356001230 [Accessed 14 Feb 2023].

Norton, H. L., Kittles, R. A., Parra, E., McKeigue, P., Mao, X., Cheng, K., Canfield, V. A., Bradley, D. G., McEvoy, B. and Shriver, M. D., 2007. Genetic Evidence for the Convergent Evolution of Light Skin in Europeans and East Asians. *Molecular Biology and Evolution*, 24 (3), 710–722.

Novotny, J. A., 2011. Molybdenum Nutriture in Humans. *Journal of Evidence-Based Complementary & Alternative Medicine*, 16 (3), 164–168.

Nyakatura, K. and Bininda-Emonds, O. R., 2012. Updating the evolutionary history of Carnivora (Mammalia): a new species-level supertree complete with divergence time estimates. *BMC Biology*, 10 (1), 12.

Ogle, R. S., Maier, K. J., Kiffney, P., Williams, M. J., Brasher, A., Melton, L. A. and Knight, A. W., 1988. Bioaccumulation of Selenium in Aquatic Ecosystems. *Lake and Reservoir Management*, 4 (2), 165–173.

Ohta, T., 1973. Slightly Deleterious Mutant Substitutions in Evolution. *Nature*, 246 (5428), 96–98.

Ohta, T., 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theoretical Population Biology*, 10 (3), 254–275.

Olivares, M., Walter, T., Hertrampf, E. and Pizarro, F., 1999. Anaemia and iron deficiency disease in children. *British Medical Bulletin*, 55 (3), 534–543.

O'Neal, S. L. and Zheng, W., 2015. Manganese Toxicity Upon Overexposure: a Decade in Review. *Current Environmental Health Reports*, 2 (3), 315–328.

Ooi, D. S. Q., Tan, V. M. H., Ong, S. G., Chan, Y. H., Heng, C. K. and Lee, Y. S., 2017. Differences in AMY1 Gene Copy Numbers Derived from Blood, Buccal Cells and Saliva Using Quantitative and Droplet Digital PCR Methods: Flagging the Pitfall. *PLoS ONE*, 12 (1), e0170767.

Orr, H. A., 2003. The distribution of fitness effects among beneficial mutations. *Genetics*, 163 (4), 1519–1526.

Osier, M. V., Pakstis, A. J., Soodyall, H., Comas, D., Goldman, D., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L. O., Bertranpetit, J., Bonne-Tamir, B., Lu, R.-B., Kidd, J. R. and Kidd, K. K., 2002. A Global Perspective on Genetic Variation at the ADH Genes Reveals Unusual Patterns of Linkage Disequilibrium and Diversity. *American Journal of Human Genetics*, 71 (1), 84–99.

Padoa, C., Goldman, A., Jenkins, T. and Ramsay, M., 1999. Cystic fibrosis carrier frequencies in populations of African origin. *Journal of Medical Genetics*, 36 (1), 41–44.

Pagani, L., Kivisild, T., Tarekegn, A., Ekong, R., Plaster, C., Gallego Romero, I., Ayub, Q., Mehdi, S. Q., Thomas, M. G., . . . Tyler-Smith, C., 2012. Ethiopian genetic diversity reveals linguistic stratification and complex influences on the Ethiopian gene pool. *American Journal of Human Genetics*, 91 (1), 83–96.

Pagani, L., Schiffels, S., Gurdasani, D., Danecek, P., Scally, A., Chen, Y., Xue, Y., Haber, M., Ekong, R., . . . Tyler-Smith, C., 2015. Tracing the Route of Modern Humans out of Africa by Using 225 Human Genome Sequences from Ethiopians and Egyptians. *American Journal of Human Genetics*, 96 (6), 986–991.

Papp, L. V., Lu, J., Holmgren, A. and Khanna, K. K., 2007. From selenium to selenoproteins: synthesis, identity, and their role in human health. *Antioxidants & Redox Signaling*, 9 (7), 775–806.

Patin, E., Laval, G., Barreiro, L. B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K. K., Kidd, J. R., Veen, L. V. der, Hombert, J.-M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E. and Quintana-Murci, L., 2009. Inferring the Demographic History of African Farmers and

Pygmy Hunter–Gatherers Using a Multilocus Resequencing Data Set. *PLOS Genetics*, 5 (4), e1000448.

Patin, E., Lopez, M., Grollemund, R., Verdu, P., Harmant, C., Quach, H., Laval, G., Perry, G. H., Barreiro, L. B., . . . Quintana-Murci, L., 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science*, 356 (6337), 543–546.

Peischl, S., Dupanloup, I., Bosshard, L. and Excoffier, L., 2016. Genetic surfing in human populations: from genes to genomes. *Current Opinion in Genetics & Development*, 41, 53–61.

Peraza, M. A., Ayala-Fierro, F., Barber, D. S., Casarez, E. and Rael, L. T., 1998. Effects of micronutrients on metal toxicity. *Environmental Health Perspectives*, 106 (Suppl 1), 203–216.

Perry, G. H. and Dominy, N. J., 2009. Evolution of the human pygmy phenotype. *Trends in Ecology & Evolution*, 24 (4), 218–225.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C. and Stone, A. C., 2007. Diet and the evolution of human amylase gene copy number variation. *Nature Genetics*, 39 (10), 1256–1260.

Peter, B. M., Huerta-Sanchez, E. and Nielsen, R., 2012. Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLOS Genetics*, 8 (10), e1003011.

Peters, M. M., Hill, K. E., Burk, R. F. and Weeber, E. J., 2006. Altered hippocampus synaptic function in selenoprotein P deficient mice. *Molecular Neurodegeneration*, 1 (1), 12.

Phillips, P. C., 2008. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9 (11), 855–867.

Piatigorsky, J. and Wistow, G., 1991. The Recruitment of Crystallins: New Functions Precede Gene Duplication. *Science*, 252 (5009), 1078–1079.

Pickrell, J. K., Patterson, N., Loh, P.-R., Lipson, M., Berger, B., Stoneking, M., Pakendorf, B. and Reich, D., 2014. Ancient west Eurasian ancestry in southern and eastern Africa. *Proceedings of the National Academy of Sciences*, 111 (7), 2632–2637.

Pierron, D., Heiske, M., Razafindrazaka, H., Pereda-Loth, V., Sanchez, J., Alva, O., Arachiche, A., Boland, A., Olaso, R., . . . Letellier, T., 2018. Strong selection during the last millennium for African ancestry in the admixed population of Madagascar. *Nature Communications*, 9 (1).

Pietrangelo, A., 2015. Pathogens, Metabolic Adaptation, and Human Diseases—An Iron-Thrifty Genetic Model. *Gastroenterology*, 149 (4), 834–838.

Pietschmann, N., Rijntjes, E., Hoeg, A., Stoedter, M., Schweizer, U., Seemann, P. and Schomburg, L., 2014. Selenoprotein P is the essential selenium transporter for bones. *Metallomics*, 6 (5), 1043–1049.

Pike, V. and Zlotkin, S., 2019. Excess micronutrient intake: defining toxic effects and upper limits in vulnerable populations. *Annals of the New York Academy of Sciences*, 1446 (1), 21–43.

Plum, L. M., Rink, L. and Haase, H., 2010. The Essential Toxin: Impact of Zinc on Human Health. *International Journal of Environmental Research and Public Health*, 7 (4), 1342–1365.

Pope, K. O. and Terrell, J. E., 2007. Environmental setting of human migrations in the circum-Pacific region. *Journal of Biogeography*, 0 (0), 071009214220006-???

Popejoy, A. B. and Fullerton, S. M., 2016. Genomics is failing on diversity. *Nature*, 538 (7624), 161–164.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R. and Finn, R. D., 2018. HMMER web server: 2018 update. *Nucleic Acids Research*, 46 (W1), W200–W204.
Prasad, A. S., 2013. Discovery of Human Zinc Deficiency: Its Impact on Human Health and Disease. *Advances in Nutrition*, 4 (2), 176–190.
Prezeworski, M., Coop, G. and Wall, J. D., 2005. The Signature of Positive Selection on Standing Genetic Variation. *Evolution*, 59 (11), 2312–2323.
Pritchard, J. K. and Di Rienzo, A., 2010. Adaptation – not by sweeps alone. *Nature Reviews Genetics*, 11 (10), 665–667.
Pritchard, J. K., Pickrell, J. K. and Coop, G., 2010. The Genetics of Human Adaptation: Hard Sweeps, Soft Sweeps, and Polygenic Adaptation. *Current Biology*, 20 (4), R208–R215.
Pritchard, J. K., Stephens, M. and Donnelly, P., 2000. Inference of population structure using multilocus genotype data. *Genetics*, 155 (2), 945–959.
Prohaska, J. R., 2014. Impact of copper deficiency in humans. *Annals of the New York Academy of Sciences*, 1314 (1), 1–5.
Prüfer, K., Racimo, F., Patterson, N., Jay, F., Sankararaman, S., Sawyer, S., Heinze, A., Renaud, G., Sudmant, P. H., . . . Pääbo, S., 2014. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505 (7481), 43–49.
Prugnolle, F., Manica, A. and Balloux, F., 2005. Geography predicts neutral genetic diversity of human populations. *Current biology : CB*, 15 (5), R159–R160.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., Bakker, P. I. W. de, Daly, M. J. and Sham, P. C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81 (3), 559–575.
Pybus, M., Luisi, P., Dall'Olio, G. M., Uzkudun, M., Laayouni, H., Bertranpetit, J. and Engelken, J., 2015. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. *Bioinformatics*, 31 (24), 3946–3952.
Qin, X., Chiang, C. W. K., Gaggiotti, O. 2022. Deciphering signatures of natural selection via deep learning, *Briefings in Bioinformatics*, 23 (5)
Quintana-Murci, L., Semino, O., Bandelt, H. J., Passarino, G., McElreavey, K. and Santachiara-Benerecetti, A. S., 1999. Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature Genetics*, 23 (4), 437–441.
Racimo, F., Sankararaman, S., Nielsen, R. and Huerta-Sánchez, E., 2015. Evidence for archaic adaptive introgression in humans. *Nature Reviews Genetics*, 16 (6), 359–371.
Raj, S.M., Pagani, L., Gallego Romero, I. Kivisilid, T., Amos, W. 2013. A general linear model-based approach for inferring selection to climate. *BMC Genet* **14**, 87.
Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W. and Cavalli-Sforza, L. L., 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, 102 (44), 15942–15947.
Rambaut, A. and Grassly, N. C., 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences: CABIOS*, 13 (3), 235–238.
Rasmussen, M., Anzick, S. L., Waters, M. R., Skoglund, P., DeGiorgio, M., Stafford, T. W., Rasmussen, S., Moltke, I., Albrechtsen, A., . . . Willerslev, E., 2014. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, 506 (7487), 225–229.
Rasmussen, M. D., Hubisz, M. J., Gronau, I. and Siepel, A., 2013. *Genome-wide inference of ancestral recombination graphs* [online]. arXiv.org. Available from: https://arxiv.org/abs/1306.5110v3 [Accessed 10 Feb 2023].

Rasmussen, M. D., Hubisz, M. J., Gronau, I. and Siepel, A., 2014. Genome-Wide Inference of Ancestral Recombination Graphs. *PLOS Genetics*, 10 (5), e1004342.

Rasmussen, M., Guo, X., Wang, Y., Lohmueller, K. E., Rasmussen, S., Albrechtsen, A., Skotte, L., Lindgreen, S., Metspalu, M., . . . Willerslev, E., 2011. An Aboriginal Australian genome reveals separate human dispersals into Asia. *Science (New York, N.Y.)*, 334 (6052), 94–98.

Rayman, M. P., 2012. Selenium and human health. *The Lancet*, 379 (9822), 1256–1268.

Razzaque, M. S., 2011. Phosphate toxicity: new insights into an old problem. *Clinical science (London, England : 1979)*, 120 (3), 91–97.

Redmond, J., Palla, L., Yan, L., Jarjou, L. M. A., Prentice, A. and Schoenmakers, I., 2015. Ethnic differences in urinary calcium and phosphate excretion between Gambian and British older adults. *Osteoporosis International*, 26 (3), 1125–1135.

Rees, J. and Andrés, A., 2022. Inferring human evolutionary history. *Science*, 375 (6583), 817–818.

Rees, J. S., Castellano, S. and Andrés, A. M., 2020. The Genomics of Human Local Adaptation. *Trends in Genetics*, 36 (6), 415–428.

Rees, J., Sarangi, G., Cheng, Q., Floor, M., Andrés, A. M., Miguel, B. O., Villà-Freixa, J., Arnér, E. S. and Castellano, S., 2023. Ancient loss of catalytic selenocysteine spurred convergent adaptation in a mammalian oxidoreductase. [online]. Available from: https://www.biorxiv.org/content/10.1101/2023.01.03.522577v1 [Accessed 24 Feb 2023].

Rehman, A., Farooq, M., Ullah, A., Nadeem, F., Im, S. Y., Park, S. K. and Lee, D.-J., 2020. Agronomic Biofortification of Zinc in Pakistan: Status, Benefits, and Constraints. *Frontiers in Sustainable Food Systems* [online], 4. Available from: https://www.frontiersin.org/articles/10.3389/fsufs.2020.591722 [Accessed 17 Feb 2023].

Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., . . . Pääbo, S., 2010. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, 468 (7327), 1053–1060.

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M. R., Pugach, I., Ko, A. M.-S., Ko, Y.-C., . . . Stoneking, M., 2011. Denisova Admixture and the First Modern Human Dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*, 89 (4), 516–528.

Reich, H. J. and Hondal, R. J., 2016. Why Nature Chose Selenium. *ACS Chemical Biology*, 11 (4), 821–841.

Reilly, P. F., Tjahjadi, A., Miller, S. L., Akey, J. M. and Tucci, S., 2022. The contribution of Neanderthal introgression to modern human traits. *Current Biology*, 32 (18), R970–R983.

Renaguli, A., Luo, Y., Wang, X., Dilidaer, Y., Muyeshsaer, W., Guzailinuer, J., Zhang, Y., Xin, Y. and Guo, Y., 2018. Relationship between thyrotropin and urine iodine in Han and Uygur nationalities pregnancy women in People's Hospital of Xinjiang Uygur Autonomous Region. *Zhonghua fu chan ke za zhi*, 53, 595–601.

Rennell, D., Bouvier, S. E., Hardy, L. W. and Poteete, A. R., 1991. Systematic mutation of bacteriophage T4 lysozyme. *Journal of Molecular Biology*, 222 (1), 67–88.

Renwick, A. G., 2006. Toxicology of Micronutrients: Adverse Effects and Uncertainty. *The Journal of Nutrition*, 136 (2), 493S-501S.

Reyes-Centeno, H., Ghirotto, S., Détroit, F., Grimaud-Hervé, D., Barbujani, G. and Harvati, K., 2014. Genomic and cranial phenotype data support multiple modern human dispersals from Africa and a southern route into Asia. *Proceedings of the National Academy of Sciences*, 111 (20), 7248–7253.

Reyes-Centeno, H., Hubbe, M., Hanihara, T., Stringer, C. and Harvati, K., 2015. Testing modern human out-of-Africa dispersal models and implications for modern human origins. *Journal of Human Evolution*, 87, 95–106.

Reyes-Reali, J., Mendoza-Ramos, M.I., Garrido-Guerrero, E., Méndez-Catalá, C.F., Méndez-Cruz, A.R., Pozo-Molina, G. 2018. Hypohidrotic ectodermal dysplasia: clinical and molecular review. *Int J Dermatol*. (8):965-972.

Roca-Umbert, A., Caro-Consuegra, R., Londono-Correa, D., Rodriguez-Lozano, G. F., Vicente, R. and Bosch, E., 2022. Understanding signatures of positive natural selection in human zinc transporter genes. *Scientific Reports*, 12 (1), 4320.

Romagné, F., Santesmasses, D., White, L., Sarangi, G. K., Mariotti, M., Hübler, R., Weihmann, A., Parra, G., Gladyshev, V. N., Guigó, R. and Castellano, S., 2014. SelenoDB 2.0: annotation of selenoprotein genes in animals and their genetic diversity in humans. *Nucleic Acids Research*, 42 (Database issue), D437–D443.

Rose, J. I., Usik, V. I., Marks, A. E., Hilbert, Y. H., Galletti, C. S., Parton, A., Geiling, J. M., Černý, V., Morley, M. W. and Roberts, R. G., 2011. The Nubian Complex of Dhofar, Oman: An African Middle Stone Age Industry in Southern Arabia. *PLOS ONE*, 6 (11), e28239.

Rose, S. R., 1995. Isolated Central Hypothyroidism in Short Stature. *Pediatric Research*, 38 (6), 967–973.

Rossier, B. C., Pradervand, S., Schild, L. and Hummler, E., 2002. Epithelial Sodium Channel and the Control of Sodium Balance: Interaction Between Genetic and Environmental Factors. *Annual Review of Physiology*, 64 (1), 877–897.

Rowles, A., 2023. *Why Molybdenum Is an Essential Nutrient* [online]. Healthline. Available from: https://www.healthline.com/nutrition/molybdenum [Accessed 19 Mar 2023].

Ryan, J., Rashid, A., Torrent, J., Yau, S. K., Ibrikci, H., Sommer, R. and Erenoglu, E. B., 2013. Chapter One - Micronutrient Constraints to Crop Production in the Middle East–West Asia Region: Significance, Research, and Management. *In*: Sparks, D. L., ed. *Advances in Agronomy* [online]. Academic Press, 1–84. Available from: https://www.sciencedirect.com/science/article/pii/B9780124171879000012 [Accessed 15 Feb 2023].

Ryan, J. and Stroehlein', J. L., n.d. Lise of Sulfuric Acid on Phorphorus Deficient Arizona Soils.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., . . . Lander, E. S., 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419 (6909), 832–837.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D. and Lander, E. S., 2006. Positive Natural Selection in the Human Lineage. *Science*, 312 (5780), 1614–1620.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F. and Lander, E. S., 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449 (7164), 913–918

Sanchez, T., Cury, J., Charpiat, G., Jay, F. 2020. Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation. *Molecular Ecology Resources.* 21(8), 2645-2660.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S. and Reich, D., 2012. The Date of Interbreeding between Neandertals and Modern Humans. *PLOS Genetics*, 8 (10), e1002947.

Santesmasses, D., Mariotti, M. and Gladyshev, V. N., 2020. Tolerance to Selenoprotein Loss Differs between Human and Mouse. *Molecular Biology and Evolution*, 37 (2), 341–354.

Sarangi, G. K., Romagné, F. and Castellano, S., 2018. Distinct Patterns of Selection in Selenium-Dependent Genes between Land and Aquatic Vertebrates. *Molecular Biology and Evolution*, 35 (7), 1744–1756.

Sarangi, G. K., White, L. and Castellano, S., 2017. Genetic Adaptation and Selenium Uptake in Vertebrates. *In*: John Wiley & Sons, Ltd, ed. *eLS* [online]. Wiley, 1–8. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0026518 [Accessed 14 Mar 2023].

Savas, S., Briollais, L., Ibrahim-zada, I., Jarjanazi, H., Choi, Y. H., Musquera, M., Fleshner, N., Venkateswaran, V. and Ozcelik, H., 2010. A Whole-Genome SNP Association Study of NCI60 Cell Line Panel Indicates a Role of Ca2+ Signaling in Selenium Resistance. *PLOS ONE*, 5 (9), e12601.

Savolainen, O., Lascoux, M. and Merilä, J., 2013. Ecological genomics of local adaptation. *Nature Reviews Genetics*, 14 (11), 807–820.

Schaefer, N. K., Shapiro, B. and Green, R. E., 2021. An ancestral recombination graph of human, Neanderthal, and Denisovan genomes. *Science Advances*, 7 (29), eabc0776.

Scherer, M. K., Trendelkamp-Schroer, B., Paul, F., Pérez-Hernández, G., Hoffmann, M., Plattner, N., Wehmeyer, C., Prinz, J.-H. and Noé, F., 2015. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *Journal of Chemical Theory and Computation*, 11 (11), 5525–5542.

Schlebusch, C. M., Gattepaille, L. M., Engström, K., Vahter, M., Jakobsson, M. and Broberg, K., 2015. Human Adaptation to Arsenic-Rich Environments. *Molecular Biology and Evolution*, 32 (6), 1544–1555.

Schlebusch, C. M., Skoglund, P., Sjödin, P., Gattepaille, L. M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M. G. B., Soodyall, H. and Jakobsson, M., 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science (New York, N.Y.)*, 338 (6105), 374–379.

Schmidt, J. M., de Manuel, M., Marques-Bonet, T., Castellano, S. and Andrés, A. M., 2019. The impact of genetic adaptation on chimpanzee subspecies differentiation. *PLoS Genetics*, 15 (11), e1008485.

Schrider, D. R. and Kern, A. D., 2016. S/HIC: Robust Identification of Soft and Hard Sweeps Using Machine Learning. *PLOS Genetics*, 12 (3), e1005928.

Schrider, D. R. and Kern, A. D., 2017. Soft Sweeps Are the Dominant Mode of Adaptation in the Human Genome. *Molecular Biology and Evolution*, 34 (8), 1863–1877.

Schrider, D. R. and Kern, A. D., 2018. Supervised Machine Learning for Population Genetics: A New Paradigm. *Trends in Genetics*, 34 (4), 301–312.

Schwarz, G., 2005. Molybdenum cofactor biosynthesis and deficiency. *Cellular and Molecular Life Sciences CMLS*, 62 (23), 2792–2810.

Secolin, R., Mas-Sandoval, A., Arauna, L. R., Torres, F. R., de Araujo, T. K., Santos, M. L., Rocha, C. S., Carvalho, B. S., Cendes, F., Lopes-Cendes, I. and Comas, D., 2019. Distribution of local ancestry and evidence of adaptation in admixed populations. *Scientific Reports*, 9 (1), 13900.

Seguin-Orlando, A., Korneliussen, T. S., Sikora, M., Malaspinas, A.-S., Manica, A., Moltke, I., Albrechtsen, A., Ko, A., Margaryan, A., Moiseyev, V., . . . Willerslev, E., 2014. Genomic structure in Europeans dating back at least 36,200 years. *Science*, 346 (6213), 1113–1118.

Serbanovic-Canic, J., Cvejic, A., Soranzo, N., Stemple, D. L., Ouwehand, W. H. and Freson, K., 2011. Silencing of RhoA nucleotide exchange factor, ARHGEF3, reveals its unexpected role in iron uptake. *Blood*, 118 (18), 4967.

Serdar, C. C., Cihan, M., Yücel, D. and Serdar, M. A., 2021. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica*, 31 (1), 010502.

Sha, Q., Pearson, W., Burcea, L. C., Wigfall, D. A., Schlesinger, P. H., Nichols, C. G. and Mercer, R. W., 2008. Human FXYD2 G41R mutation responsible for renal hypomagnesemia behaves as an inward-rectifying cation channel. *American Journal of Physiology. Renal Physiology*, 295 (1), F91–F99.

Shah, P., McCandlish, D. M. and Plotkin, J. B., 2015. Contingency and entrenchment in protein evolution under purifying selection. *Proceedings of the National Academy of Sciences*, 112 (25), E3226–E3235.

Shah, Y. M. and Xie, L., 2014. Hypoxia-Inducible Factors Link Iron Homeostasis and Erythropoiesis. *Gastroenterology*, 146 (3), 630–642.

Shahid, S. A., Zaman, M. and Heng, L., 2018. Soil Salinity: Historical Perspectives and a World Overview of the Problem. *In*: Zaman, M., Shahid, S. A., and Heng, L., eds. *Guideline for Salinity Assessment, Mitigation and Adaptation Using Nuclear and Related Techniques* [online]. Cham: Springer International Publishing, 43–53. Available from: https://doi.org/10.1007/978-3-319-96190-3_2 [Accessed 15 Feb 2023].

Sharir-Ivry, A. and Xia, Y., 2021. Quantifying evolutionary importance of protein sites: A Tale of two measures. *PLOS Genetics*, 17 (4), e1009476.

Sheehan, S., Song, Y. S. 2016. Deep learning for population inference. 2016. *PLoS Comp. Bio.* 12(3).

Shema, R., Kulicke, R., Cowley, G. S., Stein, R., Root, D. E. and Heiman, M., 2015. Synthetic lethal screening in the mammalian central nervous system identifies Gpx6 as a modulator of Huntington's disease. *Proceedings of the National Academy of Sciences*, 112 (1), 268–272.

Shenkin, A., 2006. Micronutrients in health and disease. *Postgraduate Medical Journal*, 82 (971), 559–567.

Shetaya, W. H., Young, S. D., Watts, M. J., Ander, E. L. and Bailey, E. H., 2012. Iodine dynamics in soils. *Geochimica et Cosmochimica Acta*, 77, 457–473.

Shi, W., Ayub, Q., Vermeulen, M., Shao, R., Zuniga, S., van der Gaag, K., de Knijff, P., Kayser, M., Xue, Y. and Tyler-Smith, C., 2010. A worldwide survey of human male demographic history based on Y-SNP and Y-STR data from the HGDP-CEPH populations. *Molecular Biology and Evolution*, 27 (2), 385–393.

Shi, Y., Yang, W., Tang, X., Yan, Q., Cai, X. and Wu, F., 2021. Keshan Disease: A Potentially Fatal Endemic Cardiomyopathy in Remote Mountains of China. *Frontiers in Pediatrics*, 9, 576916.

Shlisky, J., Mandlik, R., Askari, S., Abrams, S., Belizan, J. M., Bourassa, M. W., Cormick, G., Driller-Colangelo, A., Gomes, F., . . . Weaver, C., 2022. Calcium deficiency worldwide: prevalence of inadequate intakes and associated health outcomes. *Annals of the New York Academy of Sciences*, 1512 (1), 10–28.

Shukla, A. K., Dwivedi, B. S., Singh, V. K. and Gill, M. S., 2009. Macro role of micronutrients. *Indian Journal of Fertilisers*, 5 (5), 11–30.

Sillanpaeae, M., 1982. *Micronutrients and the nutrient status of soils: A global study* [online]. Rome (Italy) FAO. Available from: https://scholar.google.com/scholar_lookup?title=Micronutrients+and+the+nutrient+status+of+soils%3A+A+global+study&author=Sillanpaeae%2C+M.&publication_year=1982 [Accessed 16 Mar 2023].

da Silva Ribeiro, T., Galván, J. A. and Pool, J. E., 2022. Maximum SNP FST Outperforms Full-Window Statistics for Detecting Soft Sweeps in Local Adaptation. *Genome Biology and Evolution*, 14 (10), evac143.

Silvertooth, J., Norton, E. and Galadima, A., 2001. Evaluation of Potassium and Phosphorus Fertility In Arizona Soils.

Singh, M. V., 2009. Micronutrient nutritional problems in soils of India and improvement for human and animal health. *Indian Journal of Fertilisers*, 5 (4), 11–56.

Sirugo, G., Williams, S. M. and Tishkoff, S. A., 2019. The Missing Diversity in Human Genetic Studies. *Cell*, 177 (1), 26–31.

Skoglund, P. and Jakobsson, M., 2011. Archaic human ancestry in East Asia. *Proceedings of the National Academy of Sciences*, 108 (45), 18301–18306.

Skoglund, P., Mallick, S., Bortolini, M. C., Chennagiri, N., Hünemeier, T., Petzl-Erler, M. L., Salzano, F. M., Patterson, N. and Reich, D., 2015. Genetic evidence for two founding populations of the Americas. *Nature*, 525 (7567), 104–108.

Skoglund, P., Malmström, H., Raghavan, M., Storå, J., Hall, P., Willerslev, E., Gilbert, M. T. P., Götherström, A. and Jakobsson, M., 2012. Origins and Genetic Legacy of Neolithic Farmers and Hunter-Gatherers in Europe. *Science*, 336 (6080), 466–469.

Skoglund, P. and Mathieson, I., 2018. Ancient Genomics of Modern Humans: The First Decade. *Annual Review of Genomics and Human Genetics*, 19, 381–404.

Smith, J. M. and Haigh, J., 1974. The hitch-hiking effect of a favourable gene. *Genetical Research*, 23 (1), 23–35.

Snider, G. W., Ruggles, E., Khan, N. and Hondal, R. J., 2013. Selenocysteine Confers Resistance to Inactivation by Oxidation in Thioredoxin Reductase: Comparison of Selenium and Sulfur Enzymes. *Biochemistry*, 52 (32), 5472–5481.

Soares, P., Alshamali, F., Pereira, J. B., Fernandes, V., Silva, N. M., Afonso, C., Costa, M. D., Musilová, E., Macaulay, V., Richards, M. B., Černý, V. and Pereira, L., 2012. The Expansion of mtDNA Haplogroup L3 within and out of Africa. *Molecular Biology and Evolution*, 29 (3), 915–927.

Sohail, M., Maier, R. M., Ganna, A., Bloemendal, A., Martin, A. R., Turchin, M. C., Chiang, C. W. K., Hirschhorn, J., Daly, . . . Sunyaev, S. R., 2019. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. *eLife*, 8.

Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. and Smoller, J. W., 2013. Pleiotropy in complex traits: challenges and strategies. *Nature Reviews Genetics*, 14 (7), 483–495.

Sordillo, L. M., 2013. Selenium-Dependent Regulation of Oxidative Stress and Immunity in Periparturient Dairy Cattle. *Veterinary Medicine International*, 2013, e154045.

Spears, J. W. and Weiss, W. P., 2008. Role of antioxidants and trace elements in health and immunity of transition dairy cows. *Veterinary Journal (London, England: 1997)*, 176 (1), 70–76.

Speidel, L., Forest, M., Shi, S. and Myers, S. R., 2019. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51 (9), 1321–1329.

Stadtman, T. C., 1974. Selenium biochemistry. *Science (New York, N.Y.)*, 183 (4128), 915–922.

Stadtman, T. C., 1996. Selenocysteine. *Annual Review of Biochemistry*, 65 (1), 83–100.

Stafford, N., Wilson, C., Oceandy, D., Neyses, L. and Cartwright, E. J., 2017. The Plasma Membrane Calcium ATPases and Their Role as Major New Players in Human Disease. *Physiological Reviews*, 97 (3), 1089–1125.

Stauber, T. and Jentsch, T. J., 2013. Chloride in vesicular trafficking and function. *Annual Review of Physiology*, 75, 453–477.

Steppan, S. J., Adkins, R. M. and Anderson, J., 2004. Phylogeny and Divergence-Date Estimates of Rapid Radiations in Muroid Rodents Based on Multiple Nuclear Genes. *Systematic Biology*, 53 (4), 533–553.

Stern, A. J., Wilton, P. R. and Nielsen, R., 2019. An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. *PLOS Genetics*, 15 (9), e1008384.

Stevens, G. A., Beal, T., Mbuya, M. N. N., Luo, H., Neufeld, L. M., Addo, O. Y., Adu-Afarwuah, S., Alayón, S., Bhutta, Z., . . . Young, M. F., 2022. Micronutrient deficiencies among preschool-aged children and women of reproductive age worldwide: a pooled analysis of individual-level data from population-representative surveys. *The Lancet Global Health*, 10 (11), e1590–e1599.

Stevens, G. A., Paciorek, C. J., Flores-Urrutia, M. C., Borghi, E., Namaste, S., Wirth, J. P., Suchdev, P. S., Ezzati, M., Rohner, F., Flaxman, S. R. and Rogers, L. M., 2022. National, regional, and global estimates of anaemia by severity in women and children for 2000–19: a pooled analysis of population-representative data. *The Lancet Global Health*, 10 (5), e627–e639.

Stewart, C. and Pepper, M. S., 2016. Cystic fibrosis on the African continent. *Genetics in Medicine: Official Journal of the American College of Medical Genetics*, 18 (7), 653–662.

Stewart, C. and Pepper, M. S., 2017. Cystic Fibrosis in the African Diaspora. *Annals of the American Thoracic Society*, 14 (1), 1–7.

Stone, M. S., Martyn, L. and Weaver, C. M., 2016. Potassium Intake, Bioavailability, Hypertension, and Glucose Control. *Nutrients*, 8 (7), 444.

Storz, J. F., 2016. Causes of molecular convergence and parallelism in protein evolution. *Nature Reviews Genetics*, 17 (4), 239–250.

Streit, L., 2018. *Micronutrients: Types, Functions, Benefits and More* [online]. Healthline. Available from: https://www.healthline.com/nutrition/micronutrients [Accessed 19 Mar 2023].

Stringer, C. B. and Andrews, P., 1988. Genetic and Fossil Evidence for the Origin of Modern Humans. *Science*, 239 (4845), 1263–1268.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102 (43), 15545–15550.

Subramanian, S., 2016. The effects of sample size on population genomic analyses – implications for the tests of neutrality. *BMC Genomics*, 17 (1), 123.

Sugden, L.A., Atkinson, E.G., Fischer, A.P., Rong, S., Henn, B. M., Ramachandran, S. *2018.* Localization of adaptive variants in human genomes using averaged one-dependence estimation. *Nat Commun* **9**, 703.

Sun, Y., Zou, J., Ouyang, W. and Chen, K., 2020. Identification of PDE7B as a Potential Core Gene Involved in the Metastasis of Clear Cell Renal Cell Carcinoma. *Cancer Management and Research*, 12, 5701–5712.

Sunyecz, J. A., 2008. The use of calcium and vitamin D in the management of osteoporosis. *Therapeutics and Clinical Risk Management*, 4 (4), 827–836.

Sverrisdóttir, O. Ó., Timpson, A., Toombs, J., Lecoeur, C., Froguel, P., Carretero, J. M., Arsuaga Ferreras, J. L., Götherström, A. and Thomas, M. G., 2014. Direct estimates of natural selection in Iberia indicate calcium absorption was not the only driver of lactase persistence in Europe. *Molecular Biology and Evolution*, 31 (4), 975–983.

Swanson, E. M., Espeset, A., Mikati, I., Bolduc, I., Kulhanek, R., White, W. A., Kenzie, S. and Snell-Rood, E. C., 2016. Nutrition shapes life-history evolution across species. *Proceedings of the Royal Society B: Biological Sciences*, 283 (1834), 20152764.

Szpiech, Z. A. and Hernandez, R. D., 2014. selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. *Molecular Biology and Evolution*, 31 (10), 2824–2827.

Szpiech, Z. A., Novak, T. E., Bailey, N. P. and Stevison, L. S., 2021. Application of a novel haplotype-based scan for local adaptation to study high-altitude adaptation in rhesus macaques. *Evolution Letters*, 5 (4), 408–421.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123 (3), 585–595.

Tako, E., 2019. Dietary Trace Minerals. *Nutrients*, 11 (11), 2823.

Tang, K., Thornton, K. R. and Stoneking, M., 2007. A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. *PLOS Biology*, 5 (7), e171.

Tassi, F., Ghirotto, S., Mezzavilla, M., Vilaça, S. T., De Santi, L. and Barbujani, G., 2015. Early modern human dispersal from Africa: genomic evidence for multiple waves of migration. *Investigative Genetics*, 6, 13.

Taylor, A., Robson, A., Houghton, B. C., Jepson, C. A., Ford, W. C. L. and Frayne, J., 2013. Epididymal specific, selenium-independent GPX5 protects cells from oxidative stress-induced lipid peroxidation and DNA mutation. *Human Reproduction (Oxford, England)*, 28 (9), 2332–2342.

The UniProt Consortium, 2023. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51 (D1), D523–D531.

Thomson, C. D., 2004. Selenium and iodine intakes and status in New Zealand and Australia. *The British Journal of Nutrition*, 91 (5), 661–672.

Tian, R., Geng, Y., Yang, Y., Seim, I. and Yang, G., 2021. Oxidative stress drives divergent evolution of the glutathione peroxidase (GPX) gene family in mammals. *Integrative Zoology*, 16 (5), 696–711.

Tiffin, P. and Ross-Ibarra, J., 2014. Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology & Evolution*, 29 (12), 673–680.

Tintle, N. L., Borchers, B., Brown, M. and Bekmetjev, A., 2009. Comparing gene set analysis methods on single-nucleotide polymorphism data from Genetic Analysis Workshop 16. *BMC proceedings*, 3 Suppl 7 (Suppl 7), S96.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., . . . Williams, S. M., 2009. The Genetic Structure and History of Africans and African Americans. *Science*, 324 (5930), 1035–1044.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., . . . Deloukas, P., 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nature Genetics*, 39 (1), 31–40.

Toppo, S., Vanin, S., Bosello, V. and Tosatto, S. C. E., 2008. Evolutionary and structural insights into the multifaceted glutathione peroxidase (Gpx) superfamily. *Antioxidants & Redox Signaling*, 10 (9), 1501–1514.

Torada, L., Lorenzon, L., Beddis, A., Isildak, U., Pattini, K., Mathieson, S., Fumagalli, M. *2019.* ImaGene: a convolutional neural network to quantify natural selection from genomic data. *BMC Bioinformatics* **20** (Suppl 9), 337.

Tosatto, S. C. E., Bosello, V., Fogolari, F., Mauri, P., Roveri, A., Toppo, S., Flohé, L., Ursini, F. and Maiorino, M., 2008. The catalytic site of glutathione peroxidases. *Antioxidants & Redox Signaling*, 10 (9), 1515–1526.

Trenz, T. S., Delaix, C. L., Turchetto-Zolet, A. C., Zamocky, M., Lazzarotto, F. and Margis-Pinheiro, M., 2021. Going Forward and Back: The Complex Evolutionary History of the GPx. *Biology*, 10 (11), 1165.

Triggiani, V., Tafaro, E., Giagulli, V. A., Sabbà, C., Resta, F., Licchelli, B. and Guastamacchia, E., 2009. Role of iodine, selenium and other micronutrients in thyroid function and disorders. *Endocrine, Metabolic & Immune Disorders Drug Targets*, 9 (3), 277–294.

Trindade, S., Perfeito, L. and Gordo, I., 2010. Rate and effects of spontaneous mutations that affect fitness in mutator Escherichia coli. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365 (1544), 1177–1186.

Tucci, S. and Akey, J. M., 2019. The long walk to African genomics. *Genome Biology*, 20 (1), 130.

Tulchinsky, T. H., 2010. Micronutrient Deficiency Conditions: Global Health Issues. *Public Health Reviews*, 32 (1), 243–255.

Turchin, M. C., Chiang, C. W., Palmer, C. D., Sankararaman, S., Reich, D. and Hirschhorn, J. N., 2012. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nature Genetics*, 44 (9), 1015–1019.

Ulijaszek, S. J., Hillman, G., Boldsen, J. L. and Henry, C. J., 1991. Human Dietary Change [and Discussion]. *Philosophical Transactions: Biological Sciences*, 334 (1270), 271–279.

Underhill, P. A. and Kivisild, T., 2007. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics*, 41, 539–564.

Venkataraman, V. V., Yegian, A. K., Wallace, I. J., Holowka, N. B., Tacey, I., Gurven, M. and Kraft, T. S., 2018. Locomotor constraints favour the evolution of the human pygmy phenotype in tropical rainforests. *Proceedings of the Royal Society B: Biological Sciences*, 285 (1890), 20181492.

Vernot, B. and Akey, J. M., 2014. Resurrecting Surviving Neandertal Lineages from Modern Human Genomes. *Science*, 343 (6174), 1017–1021.

Villanea, F. A. and Schraiber, J. G., 2019. Multiple episodes of interbreeding between Neanderthal and modern humans. *Nature Ecology & Evolution*, 3 (1), 39–44.

Voight, B. F., Kudaravalli, S., Wen, X. and Pritchard, J. K., 2006. A Map of Recent Positive Selection in the Human Genome. *PLOS Biology*, 4 (3), e72.

Vyshpolsky, F., Qadir, M., Karimov, A., Mukhamedjanov, K., Bekbaev, U., Paroda, R., Aw-Hassan, A. and Karajeh, F., 2008. Enhancing the productivity of high-magnesium soil and water resources in Central Asia through the application of phosphogypsum. *Land Degradation & Development*, 19 (1), 45–56.

Wagner, G. P. and Zhang, J., 2011. The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, 12 (3), 204–213.

Wald, N. J., 2022. Folic acid and neural tube defects: Discovery, debate and the need for policy change. *Journal of Medical Screening*, 29 (3), 138–146.

Wall, J. D., Ratan, A., Stawiski, E., Wall, J. D., Stawiski, E., Ratan, A., Kim, H. L., Kim, C., Gupta, R., . . . Peterson, A. S., 2019. Identification of African-Specific Admixture between Modern and Archaic Humans. *The American Journal of Human Genetics*, 105 (6), 1254–1261.

von Wandruszka, R., 2006. Phosphorus retention in calcareous soils and the effect of organic matter on its mobility. *Geochemical Transactions*, 7 (1), 6.

Wang, K., Mathieson, I., O'Connell, J. and Schiffels, S., 2020. Tracking human population structure through time from whole genome sequences. *PLOS Genetics*, 16 (3), e1008552.

Wang, M.-H., Okazaki, T., Kugathasan, S., Cho, J. H., Isaacs, K. L., Lewis, J. D., Smoot, D. T., Valentine, J. F., Kader, H. A., . . . , G. C., Wu, Y., Datta, L. W., Hooker, S., Dassopoulos, T., Kittles, R. A., Kao, L. W. H. and Brant, S. R., 2012. Contribution of Higher Risk Genes and

European Admixture to Crohn's Disease in African Americans. *Inflammatory Bowel Diseases*, 18 (12), 2277–2287.
Wangkumhang, P. and Hellenthal, G., 2018. Statistical methods for detecting admixture. *Current Opinion in Genetics & Development*, 53, 121–127.
Weinreich, D. M., Delaney, N. F., DePristo, M. A. and Hartl, D. L., 2006. Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins. *Science*, 312 (5770), 111–114.
Weir, B. S. and Cockerham, C. C., 1984a. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38 (6), 1358–1370.
Weir, B. S. and Cockerham, C. C., 1984b. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution; International Journal of Organic Evolution*, 38 (6), 1358–1370.
Welch, R. M. and Graham, R. D., 2005. Agriculture: the real nexus for enhancing bioavailable micronutrients in food crops. *Journal of Trace Elements in Medicine and Biology*, 18 (4), 299–307.
Wessells, K. R. and Brown, K. H., 2012. Estimating the Global Prevalence of Zinc Deficiency: Results Based on Zinc Availability in National Food Supplies and the Prevalence of Stunting. *PLOS ONE*, 7 (11), e50568.
White, L., Romagné, F., Müller, E., Erlebach, E., Weihmann, A., Parra, G., Andrés, A. M. and Castellano, S., 2015. Genetic Adaptation to Levels of Dietary Selenium in Recent Human History. *Molecular Biology and Evolution*, 32 (6), 1507–1518.
White, T. D., Asfaw, B., DeGusta, D., Gilbert, H., Richards, G. D., Suwa, G. and Clark Howell, F., 2003. Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature*, 423 (6941), 742–747.
Whitlock, M. C., Griswold, C. K. and Peters, A. D., 2003. Compensating for the meltdown: The critical effective size of a population with deleterious and compensatory mutations. *Annales Zoologici Fennici*, 40 (2), 169–183.
Williams, D. M., 1983. Copper deficiency in humans. *Seminars in hematology*, 20 (2), 118–128.
Wilson, B. A., Petrov, D. A. and Messer, P. W., 2014. Soft selective sweeps in complex demographic scenarios. *Genetics*, 198 (2), 669–684.
Winkel, L. H. E., Vriens, B., Jones, G. D., Schneider, L. S., Pilon-Smits, E. and Bañuelos, G. S., 2015. Selenium Cycling Across Soil-Plant-Atmosphere Interfaces: A Critical Review. *Nutrients*, 7 (6), 4199–4239.
Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., . . . Querengesser, L., 2007. HMDB: the Human Metabolome Database. *Nucleic Acids Research*, 35 (Database issue), D521-526.
Witt, K. E. and Huerta-Sánchez, E., 2019. Convergent evolution in human and domesticate adaptation to high-altitude environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374 (1777), 20180235.
Wohns, A. W., Wong, Y., Jeffery, B., Akbari, A., Mallick, S., Pinhasi, R., Patterson, N., Reich, D., Kelleher, J. and McVean, G., 2022. A unified genealogy of modern and ancient genomes. *Science*, 375 (6583), eabi8264.
Wu, Y., 2022. Diet evolution of carnivorous and herbivorous mammals in Laurasiatheria. *BMC Ecology and Evolution*, 22 (1), 82.
Xia, Y., Hill, K. E., Byrne, D. W., Xu, J. and Burk, R. F., 2005. Effectiveness of selenium supplements in a low-selenium area of China. *The American Journal of Clinical Nutrition*, 81 (4), 829–834.

Xu, J., Ke, Z., Xia, J., He, F. and Bao, B., 2016. Change of body height is regulated by thyroid hormone during metamorphosis in flatfishes and zebrafish. *General and Comparative Endocrinology*, 236, 9–16.

Xu, J., Wang, J. and Zhao, H., 2022. The Prevalence of Kashin-Beck Disease in China: a Systematic Review and Meta-analysis. *Biological Trace Element Research*.

Xu, Y., Shan, Y., Lin, X., Miao, Q., Lou, L., Wang, Y. and Ye, J., 2021. Global patterns in vision loss burden due to vitamin A deficiency from 1990 to 2017. *Public Health Nutrition*, 24 (17), 5786–5794.

Yang, M. A., Malaspinas, A.-S., Durand, E. Y. and Slatkin, M., 2012. Ancient Structure in Africa Unlikely to Explain Neanderthal and Non-African Genetic Similarity. *Molecular Biology and Evolution*, 29 (10), 2987–2995.

Yang, Z., 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution*, 24 (8), 1586–1591.

Yang, Z., Wang, C., Nie, Y., Sun, Y., Tian, M., Ma, Y., Zhang, Y., Yuan, Y. and Zhang, L., 2021. Investigation on spatial variability and influencing factors of drinking water iodine in Xinjiang, China. *PLOS ONE*, 16 (12), e0261015.

Yang, Z., Wong, W. S. W. and Nielsen, R., 2005. Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Molecular Biology and Evolution*, 22 (4), 1107–1118.

Yant, L. J., Ran, Q., Rao, L., Van Remmen, H., Shibatani, T., Belter, J. G., Motta, L., Richardson, A. and Prolla, T. A., 2003. The selenoprotein GPX4 is essential for mouse development and protects from radiation and oxidative damage insults. *Free Radical Biology and Medicine*, 34 (4), 496–502.

Yassin, A., Debat, V., Bastide, H., Gidaszewski, N., David, J. R. and Pool, J. E., 2016. Recurrent specialization on a toxic fruit in an island Drosophila population. *Proceedings of the National Academy of Sciences*, 113 (17), 4771–4776.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., . . . Flicek, P., 2020. Ensembl 2020. *Nucleic Acids Research*, 48 (D1), D682–D688.

Ye, K., Cao, C., Lin, X., O'Brien, K. O. and Gu, Z., 2015. Natural selection on HFE in Asian populations contributes to enhanced non-heme iron absorption. *BMC Genetics*, 16 (1), 61.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z. X. P., Pool, J. E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T. S., . . . Wang, J., 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science*, 329 (5987), 75–78.

Yudell, M., Roberts, D., DeSalle, R. and Tishkoff, S., 2016. Taking race out of human genetics. *Science*, 351 (6273), 564–565.

Zhang, C., Li, J., Tian, L., Lu, D., Yuan, K., Yuan, Y. and Xu, S., 2015. Differential Natural Selection of Human Zinc Transporter Genes between African and Non-African Populations. *Scientific Reports*, 5 (1), 9658.

Zhang, J. and Kumar, S., 1997. Detection of convergent and parallel evolution at the amino acid sequence level. *Molecular Biology and Evolution*, 14 (5), 527–536.

Zhang, X., Kim, B., Lohmueller, K. E. and Huerta-Sánchez, E., 2020. The Impact of Recessive Deleterious Variation on Signals of Adaptive Introgression in Human Populations. *Genetics*, 215 (3), 799–812.

Zhao, K., Ishida, Y., Oleksyk, T. K., Winkler, C. A. and Roca, A. L., 2012. Evidence for selection at HIV host susceptibility genes in a West Central African human population. *BMC Evolutionary Biology*, 12 (1).

Zhivotovsky, L. A., Bennett, L., Bowcock, A. M. and Feldman, M. W., 2000. Human Population Expansion and Microsatellite Variation. *Molecular Biology and Evolution*, 17 (5), 757–767.

Zhou, Z., Tran, P. Q., Breister, A. M., Liu, Y., Kieft, K., Cowley, E. S., Karaoz, U. and Anantharaman, K., 2022. METABOLIC: high-throughput profiling of microbial genomes for functional traits, metabolism, biogeochemistry, and community-scale functional networks. *Microbiome*, 10 (1), 33.

Appendices

Chapter 2: Supplementary Material

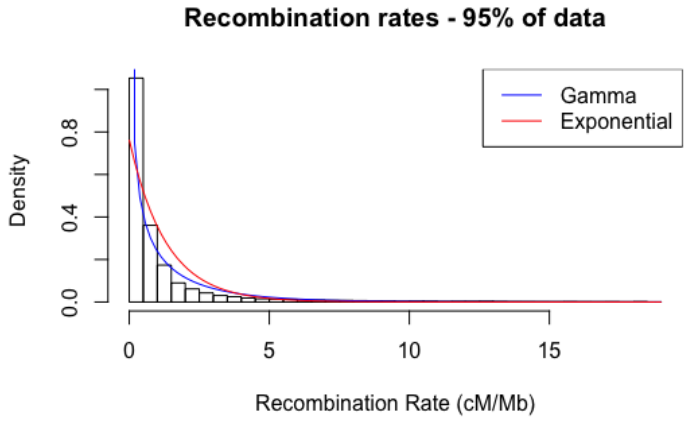


***Figure S2.1: The empirical distribution of recombination rate modelled from chromosome 15 from the HGDP dataset* (Bergström et al. 2020).** *The gamma and exponential distributions fitted (blue and red, respectively).*

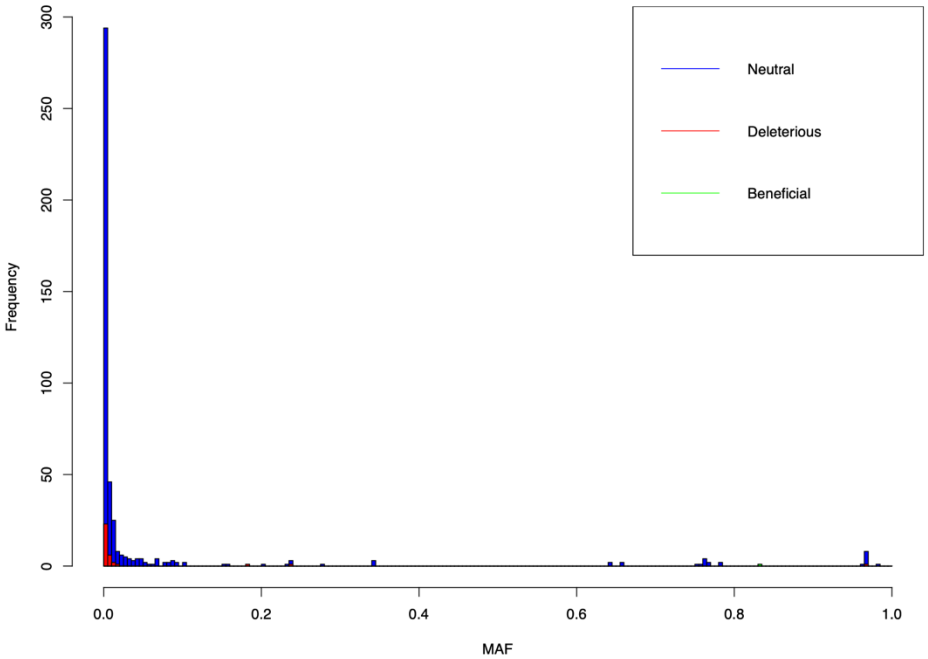


***Figure S2.2*: Example of the site frequency spectrum calculated from the VCF files given at the end of the burn-in simulation.** *Appears as expected in humans.*

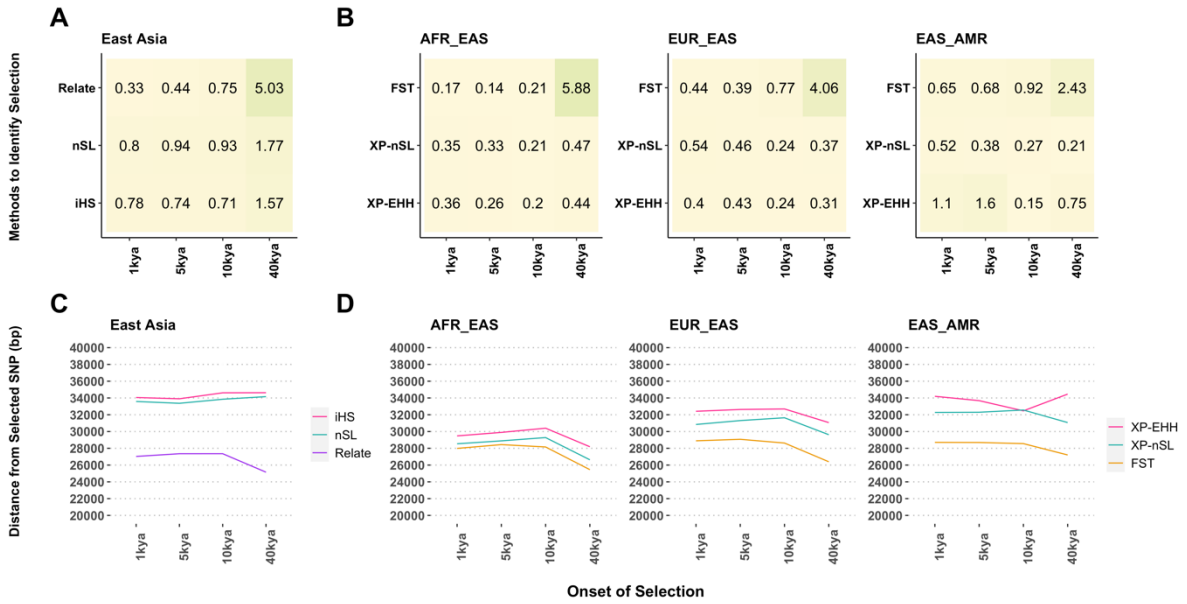


***Fig. S2.3*: Further analysis investigating selection in European populations.** *Top panel shows the percentage of tagged variants that are the SNP with the strongest evidence of selection across timepoints in the European population for A) iHS, nSL and Relate and B) the cross-population statistics XPEHH, XPnSL and F_{ST} (given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America). Bottom panel shows the average distance between the tagged variant and the top-ranking SNP for C) iHS, nSL and Relate and D) the cross-population statistics XPEHH, XPnSL and F_{ST}.*

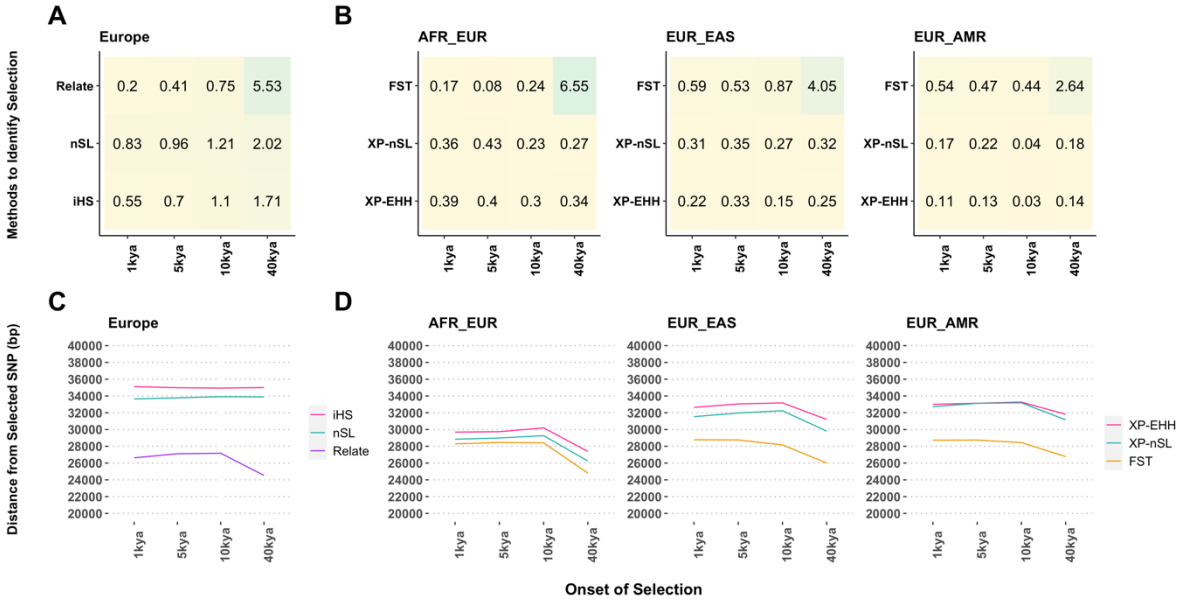


***Fig. S2.4: Further analysis investigating selection in East Asian populations.** Top panel shows the percentage of tagged variants that are the SNP with the strongest evidence of selection across timepoints in the East Asian population for A) iHS, nSL and Relate and B) the cross-population statistics XPEHH, XPnSL and F_{ST} (given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America). Bottom panel shows the average distance between the tagged variantsand the top-ranking SNP for C) iHS, nSL and Relate and D) the cross-population statistics XPEHH, XPnSL and F_{ST}.*

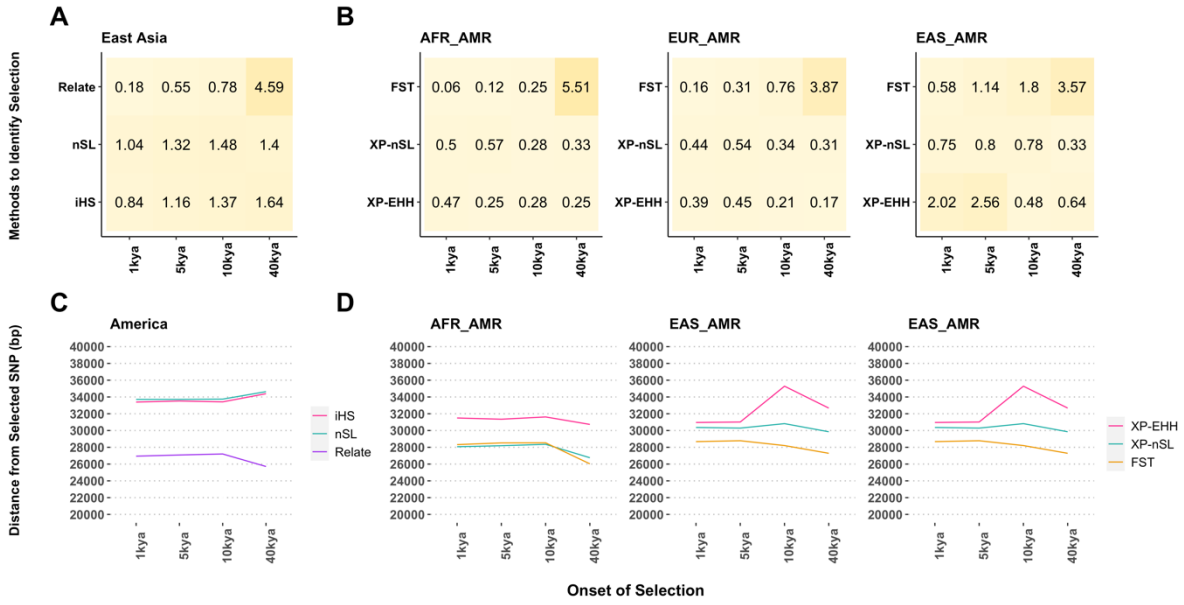


Fig. S2.5: Further analysis investigating selection in American populations. *Top panel shows the percentage of tagged variants that are the SNP with the strongest evidence of selection across timepoints in the American population for A) iHS, nSL and Relate and B) the cross-population statistics XPEHH, XPnSL and F_{ST} (given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America). Bottom panel shows the average distance between the tagged variant and the top-ranking SNP for C iHS, nSL and Relate and D) the cross-population statistics XPEHH, XPnSL and F_{ST}*

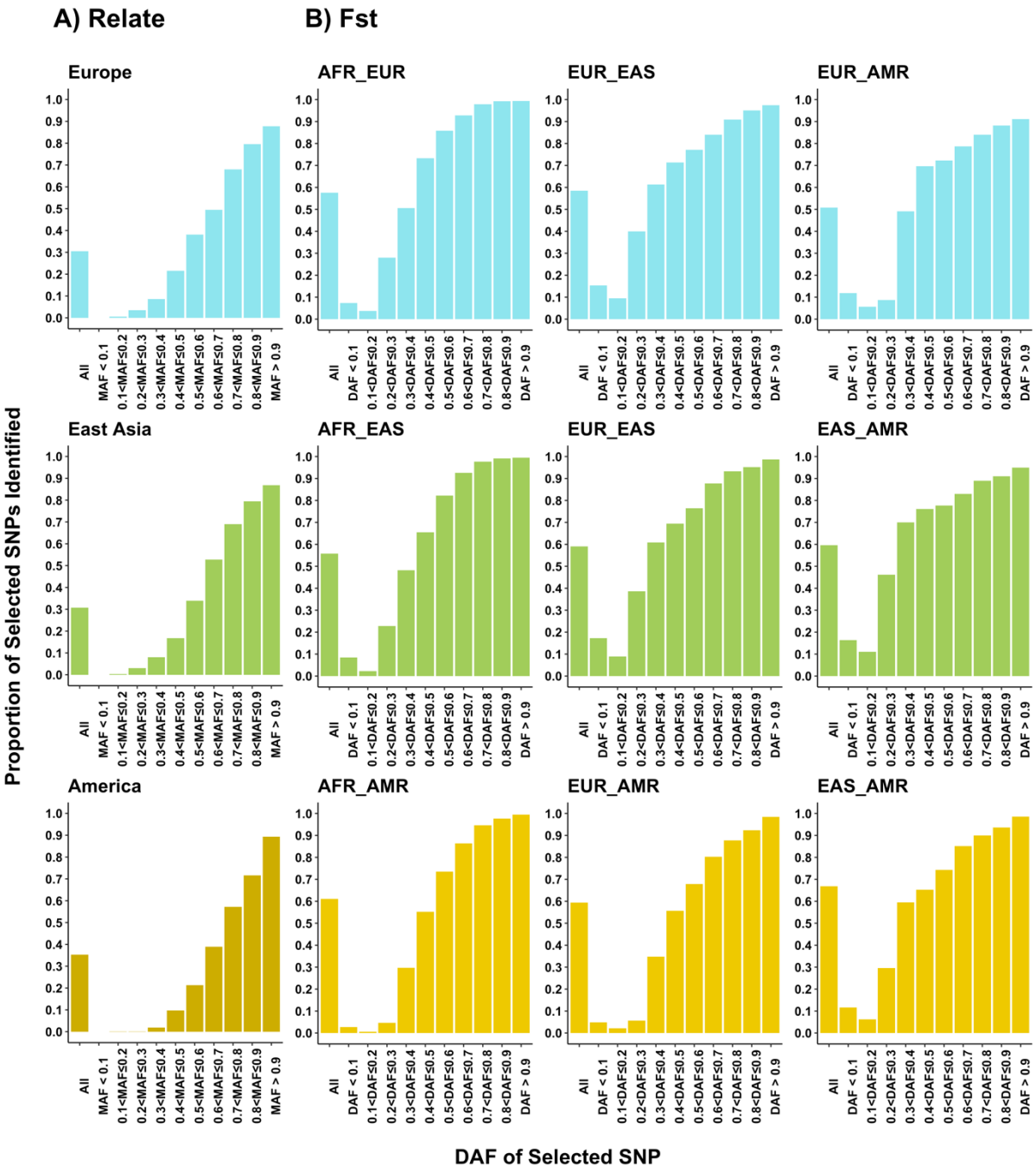


***Fig. S2.6*: The proportion of selected SNPs identified as under selection**. *Partitioned by the DAF of the tagged variant, for A) Relate and B) F_{ST} (given for three population comparisons, where AFR=Africa; EUR=Europe; EAS=East Asia; AMR=America). Given for selection acting at 40kya for the European (blue), East Asian (green) and American (yellow) populations. There are few cases of low DAF (<20%) given that the simulations condition on the tagged variant being at 10% frequency or higher, and these results may therefore be noisy at lower DAF bins.*

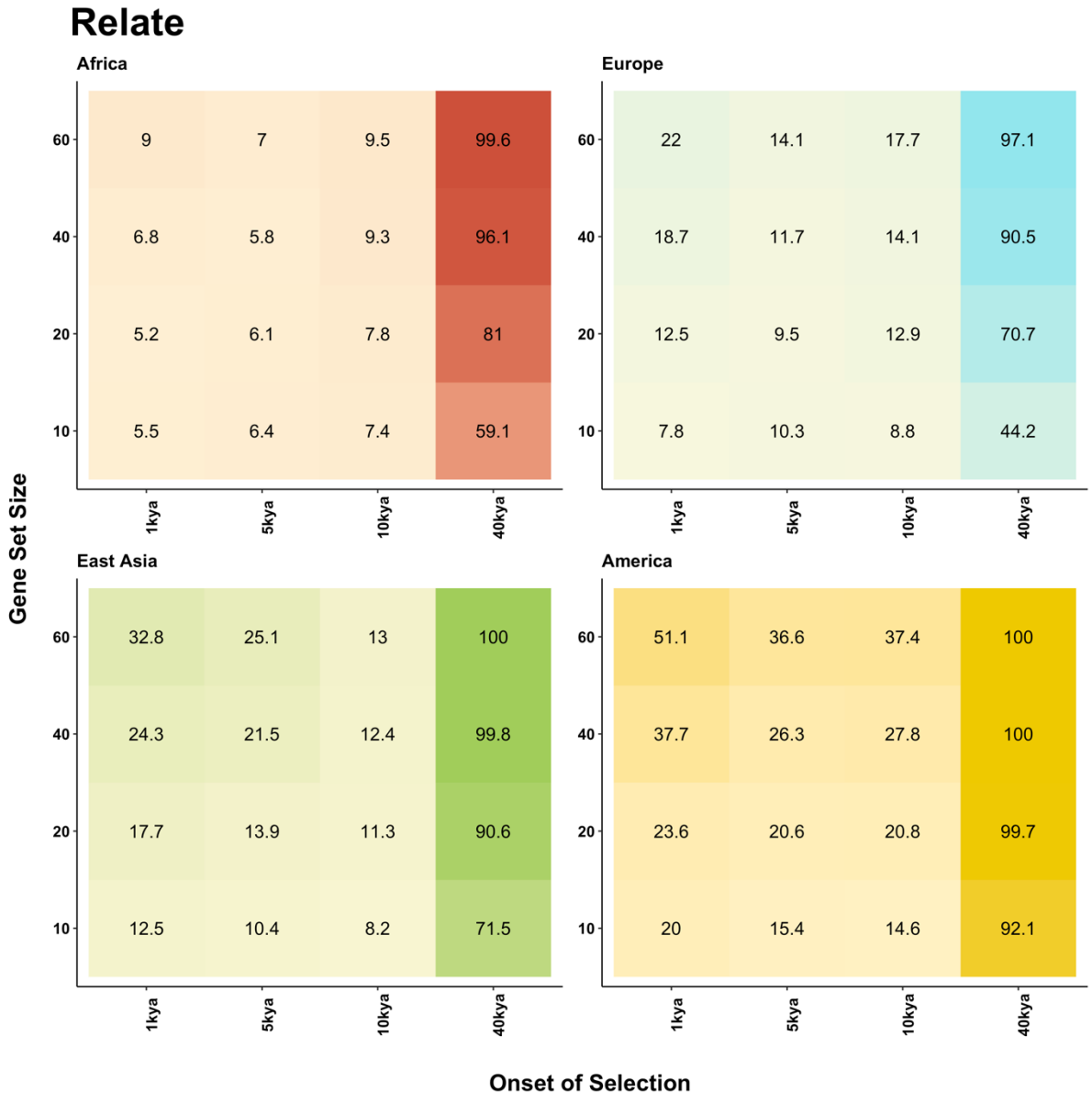


***Fig. S2.7*: The percentage of gene sets identified as being under selection according to the SUMSTAT method integrating Relate values.** *For the gene set sizes of 10, 20, 40 and 60. Shown for selection acting on four different timepoints on four different populations (as shown).*

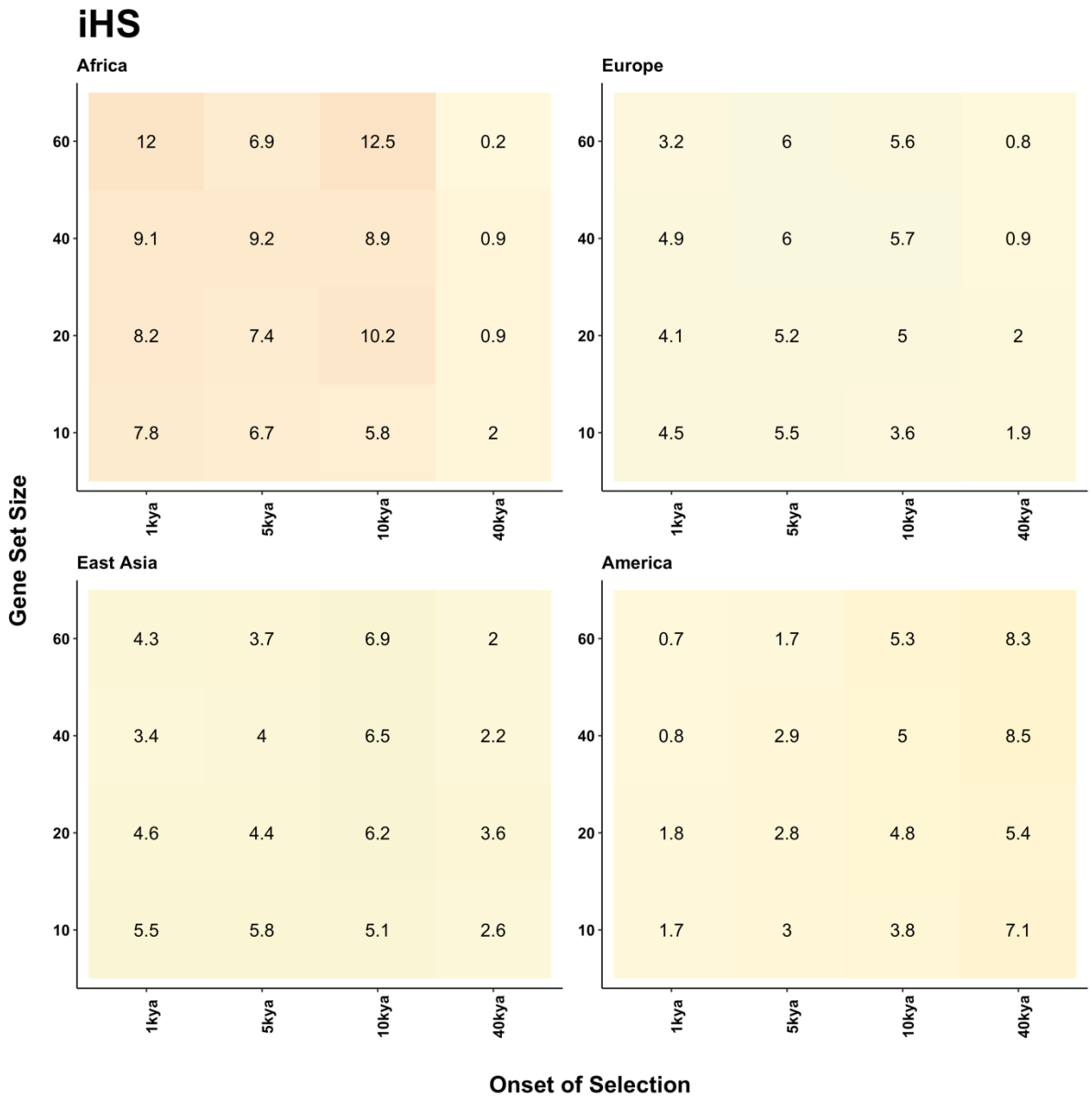


***Fig. S2.8**: **The percentage of gene sets identified as being under selection according to the SUMSTAT method integrating iHS values**. For the gene set sizes of 10, 20, 40 and 60. Shown for selection acting on four different timepoints on four different populations (as shown).*

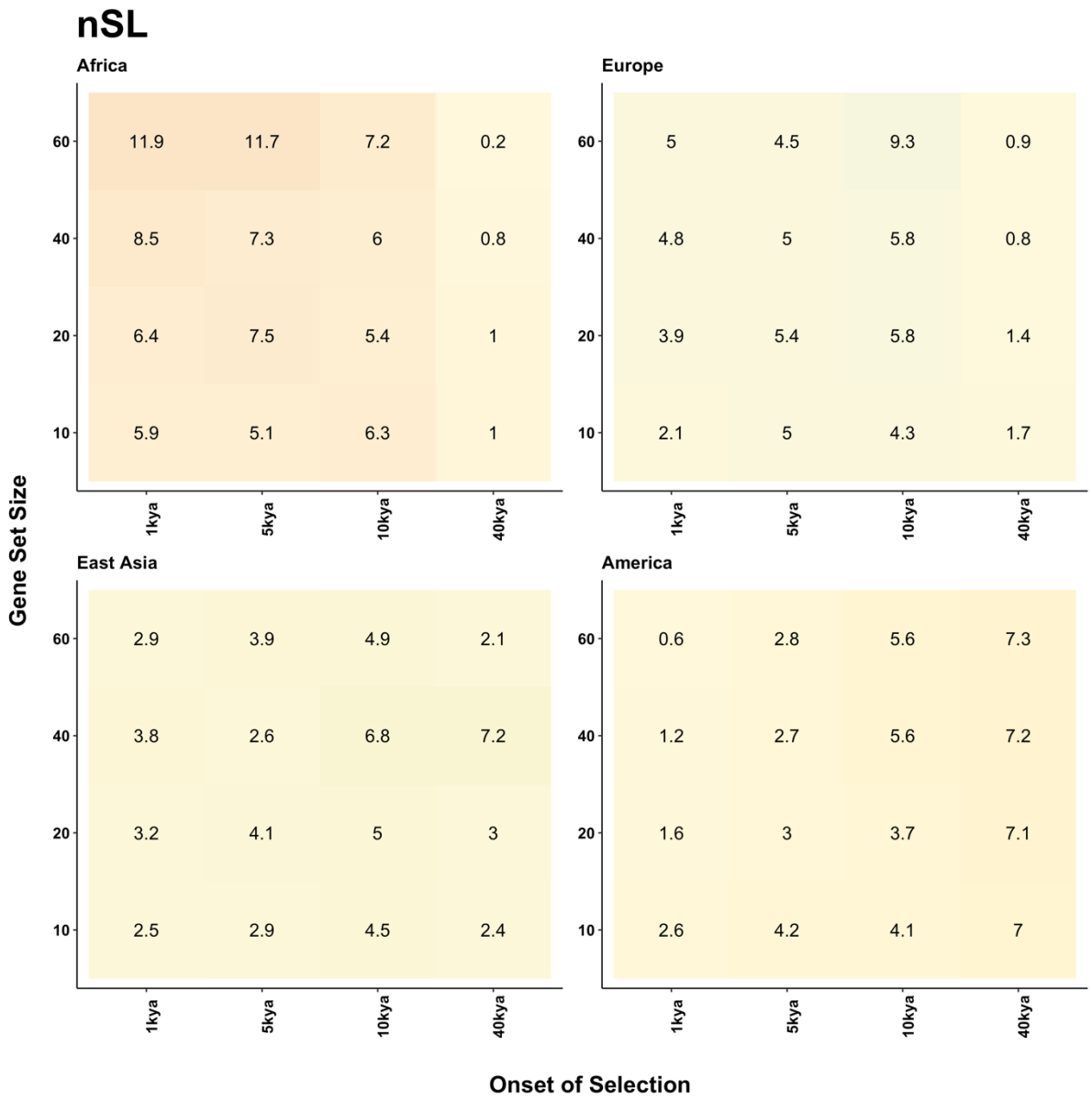


***Fig. S2.9**: **The percentage of gene sets identified as being under selection according to the SUMSTAT method integrating nSL value**. For the gene set sizes of 10, 20, 40 and 60. Shown for selection acting on four different timepoints on four different populations (as shown).*

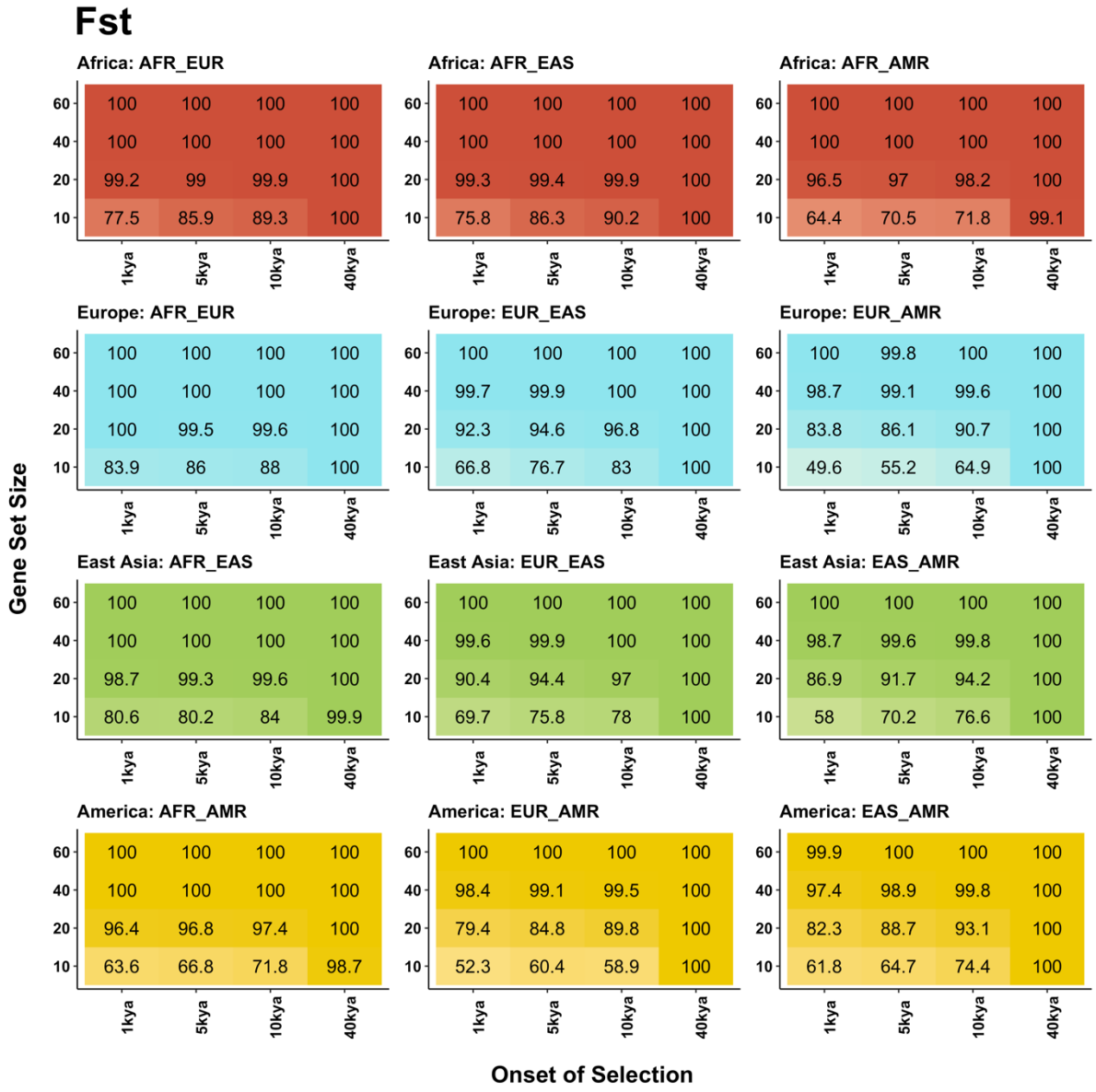


***Fig. S2.10**: The percentage of gene sets identified as being under selection according to the SUMSTAT method integrating F_{ST} values. For the gene set sizes of 10, 20, 40 and 60. Shown for selection acting on four different timepoints on four different populations for three population comparisons (given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America).*

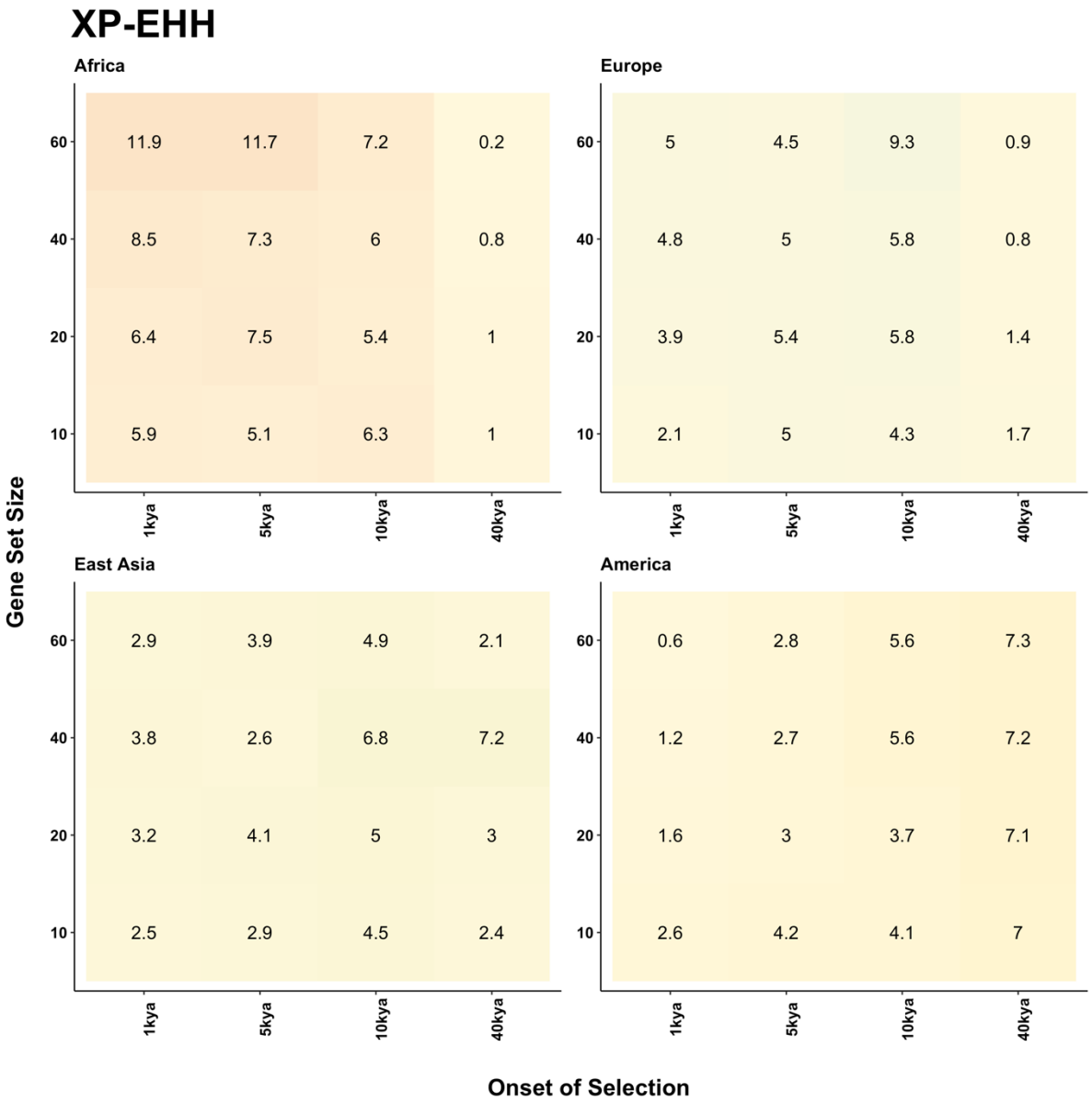


***Fig. S2.11**: **The percentage of gene sets identified as being under selection according to the SUMSTAT method integrating XPEHH values.** For the gene set sizes of 10, 20, 40 and 60. Shown for selection acting on four different timepoints on four different populations for three population comparisons (given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America).*

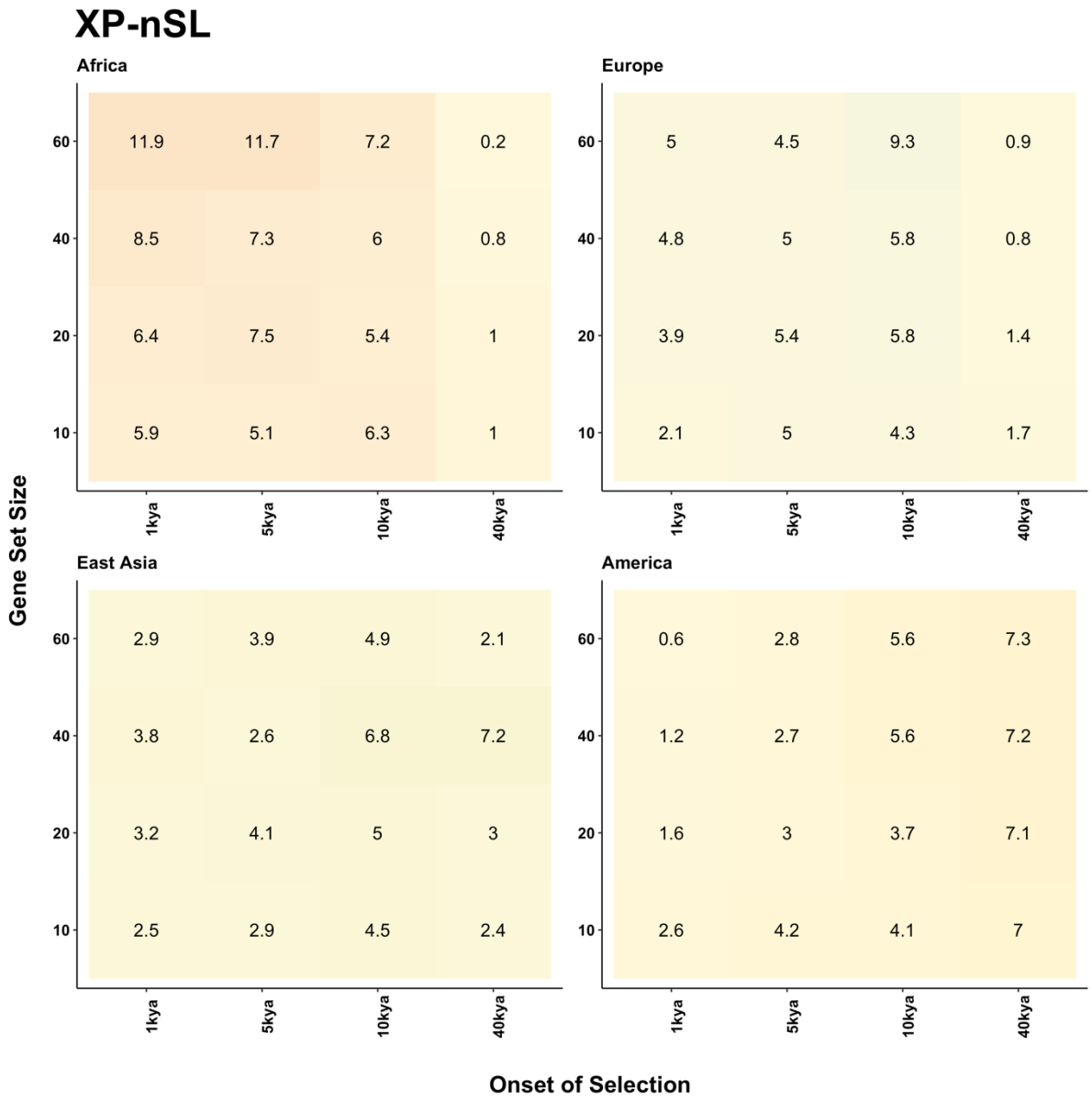


***Fig. S2.12**: **The percentage of gene sets identified as being under selection according to the SUMSTAT method integrating XPnSL values.** For the gene set sizes of 10, 20, 40 and 60. Shown for selection acting on four different timepoints on four different populations for three population comparisons (given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America).*

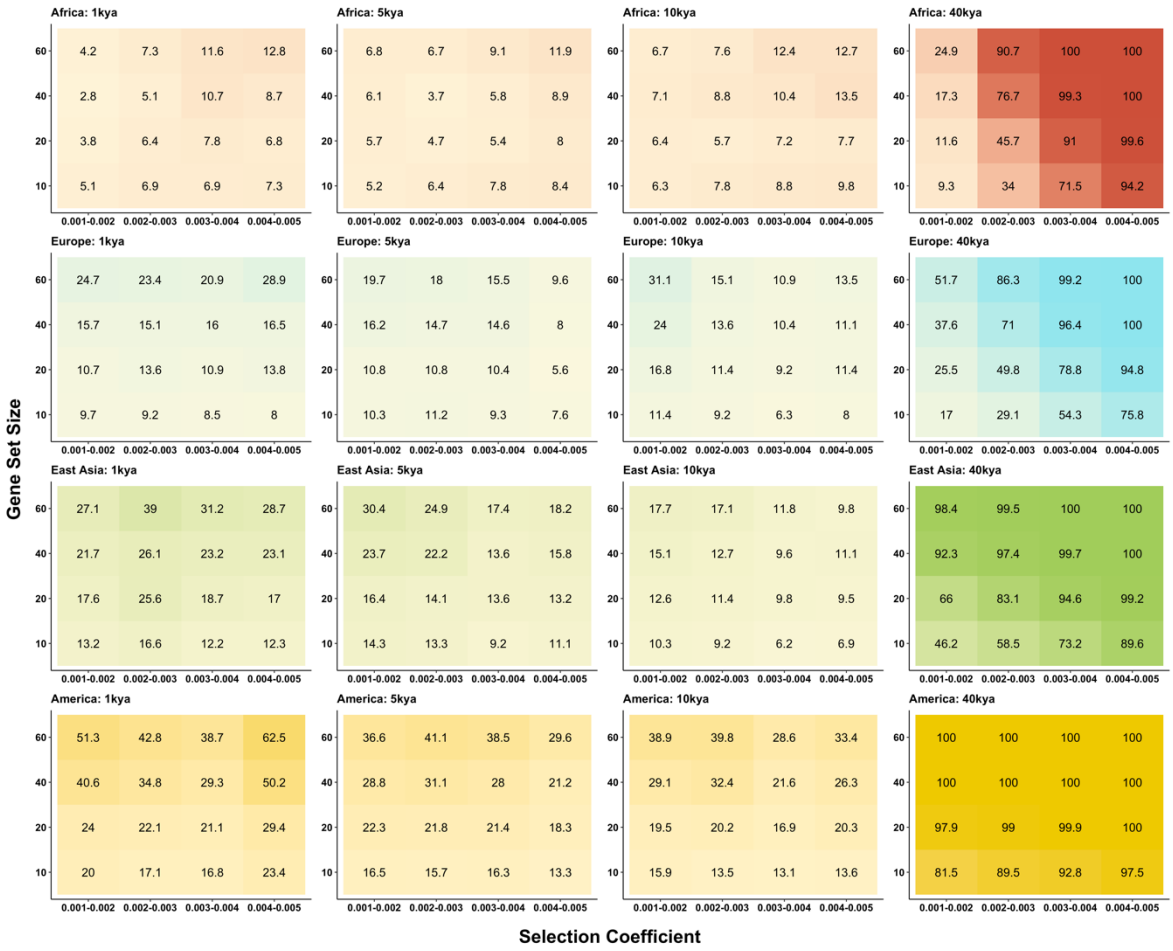


***Fig. S2.13**: The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating Relate values partitioned by selection coefficient of the tagged variant.* *Shown for selection acting on gene set sizes of 10, 20, 40 and 60, acting at 1kya, 5kya, 10kya, 40kya on four different populations.*

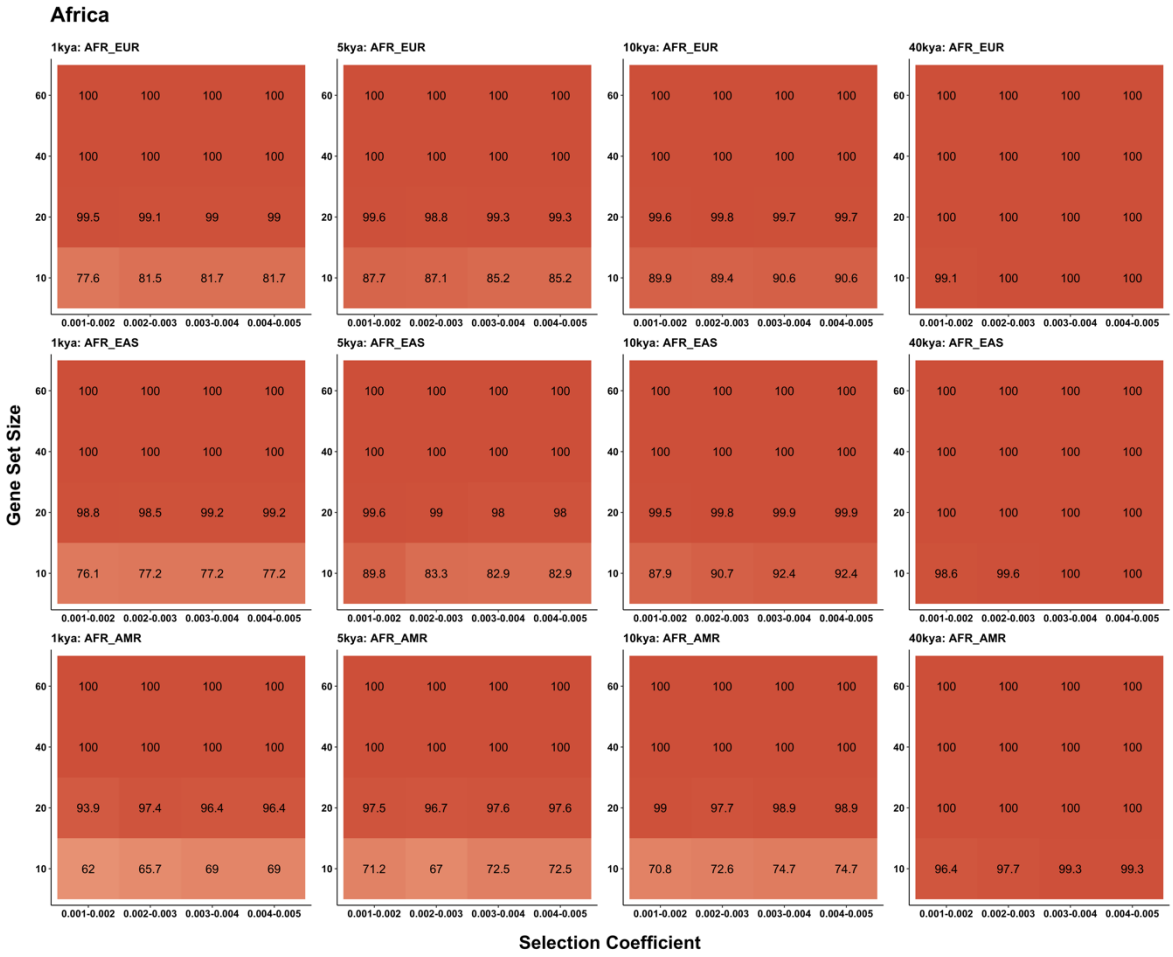


Fig. S2.14: The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating African F_{ST} values partitioned by selection coefficient of the tagged variant. *Shown for selection acting on the African population at 1kya, 5kya, 10kya, 40kya , on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

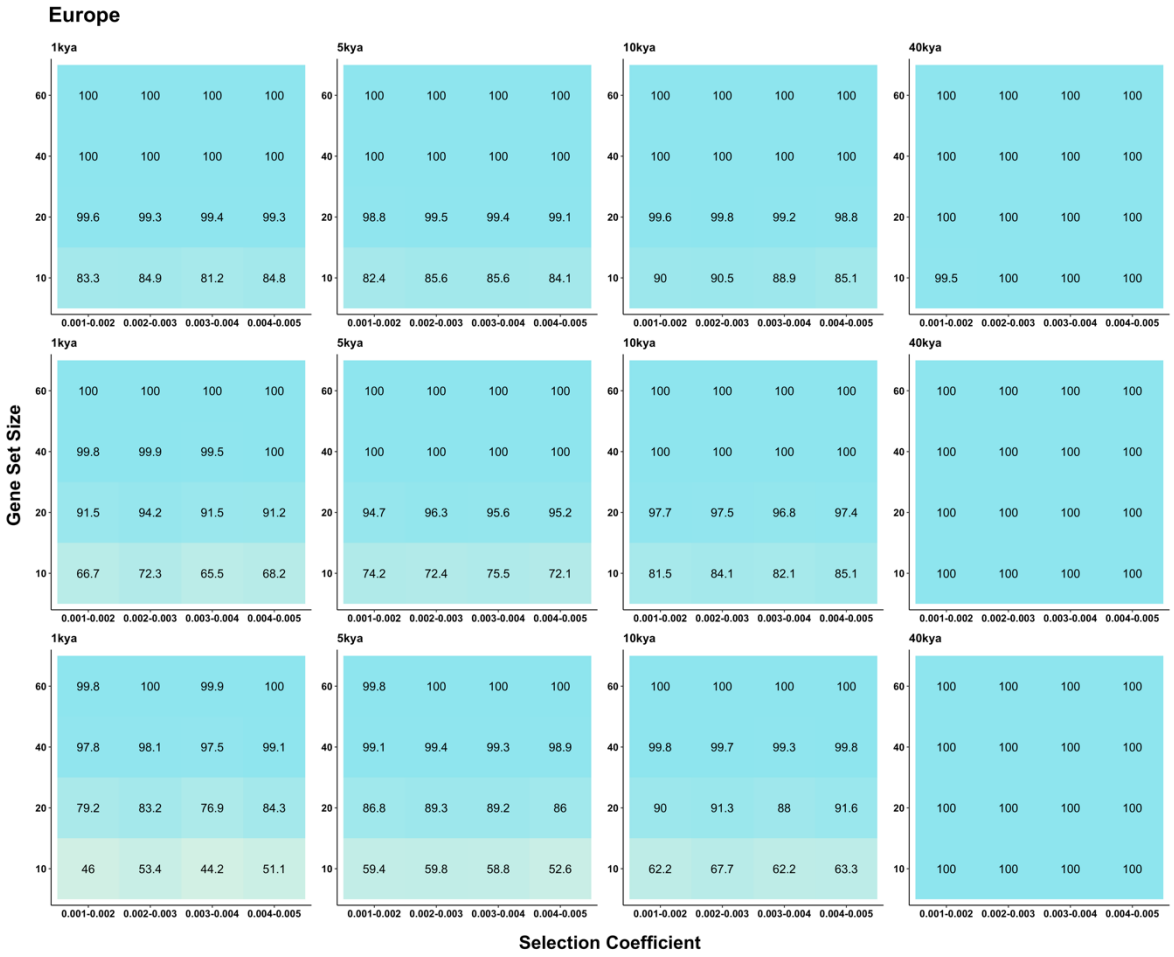


Fig. S2.15: The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating European F_{ST} values partitioned by selection coefficient of the tagged variant. *Shown for selection acting on the European population at 1kya, 5kya, 10kya, 40kya , on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

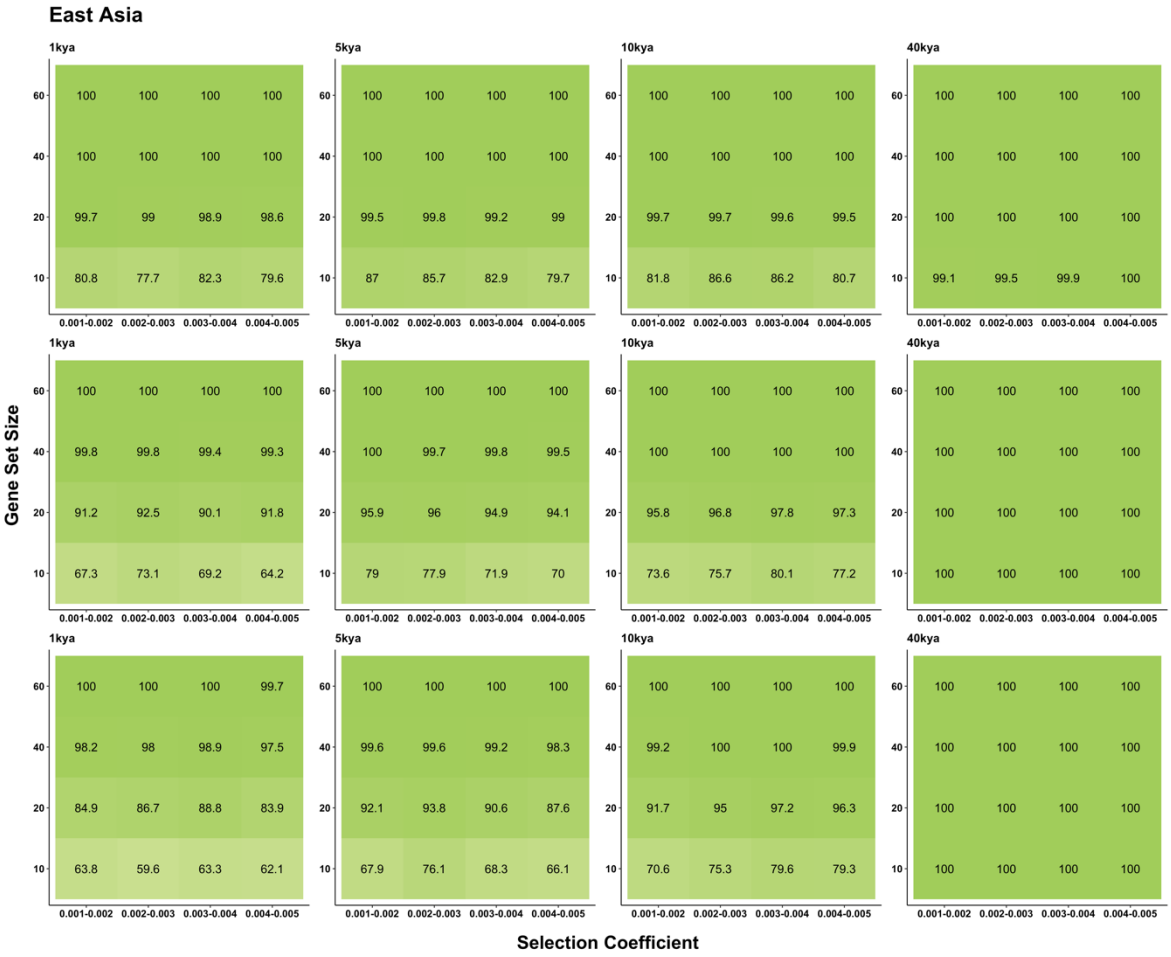


***Fig. S2.16**: **The percentage of gene sets identified as being under selection, according to SUMSTAT method integrating East Asian F_{ST} values partitioned by selection coefficient of the tagged variant.** Shown for selection acting on the East Asian population at 1kya, 5kya, 10kya, 40kya, on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

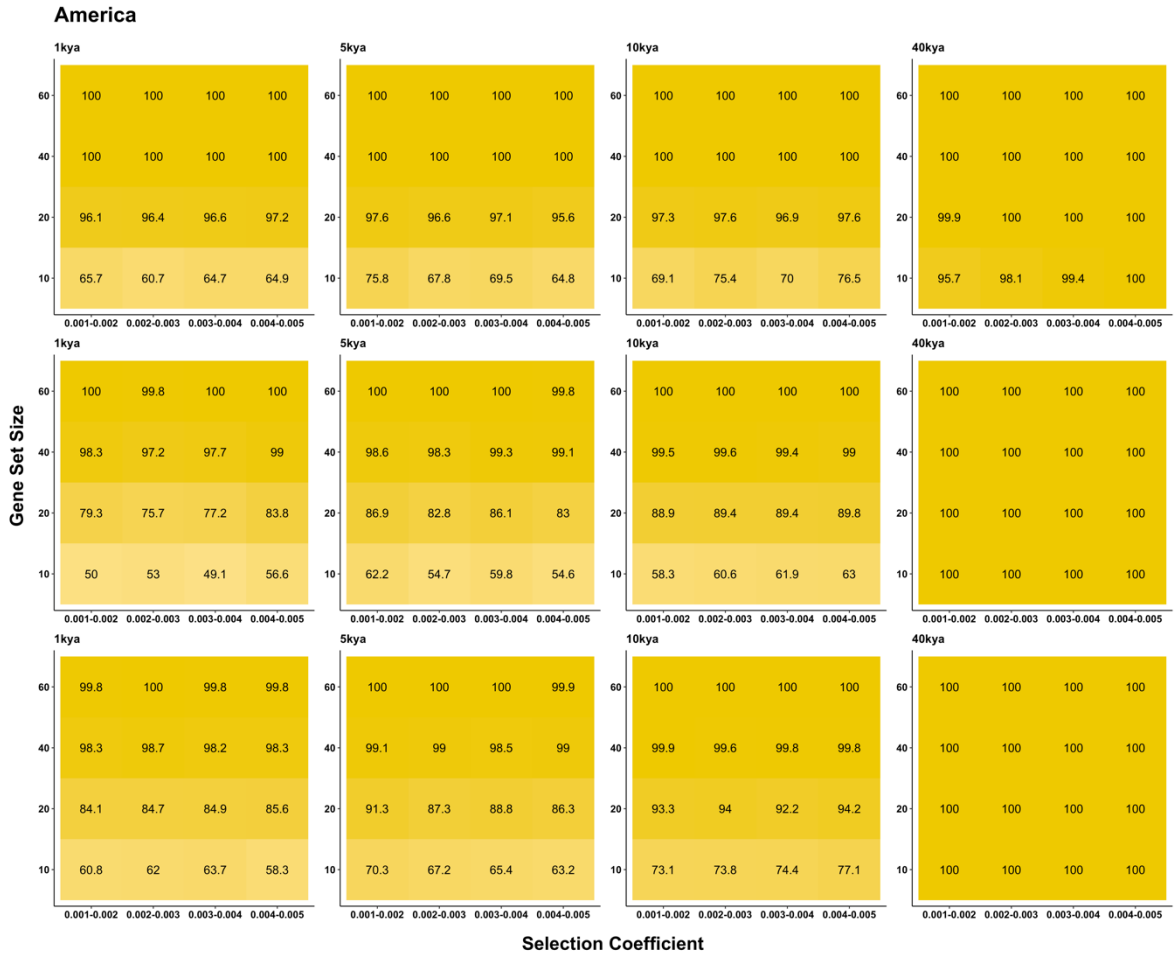


***Fig. S2.17**: **The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating American F_{ST} values partitioned by selection coefficient of the tagged variant.**. Shown for selection acting on the American population at 1kya, 5kya, 10kya, 40kya, on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

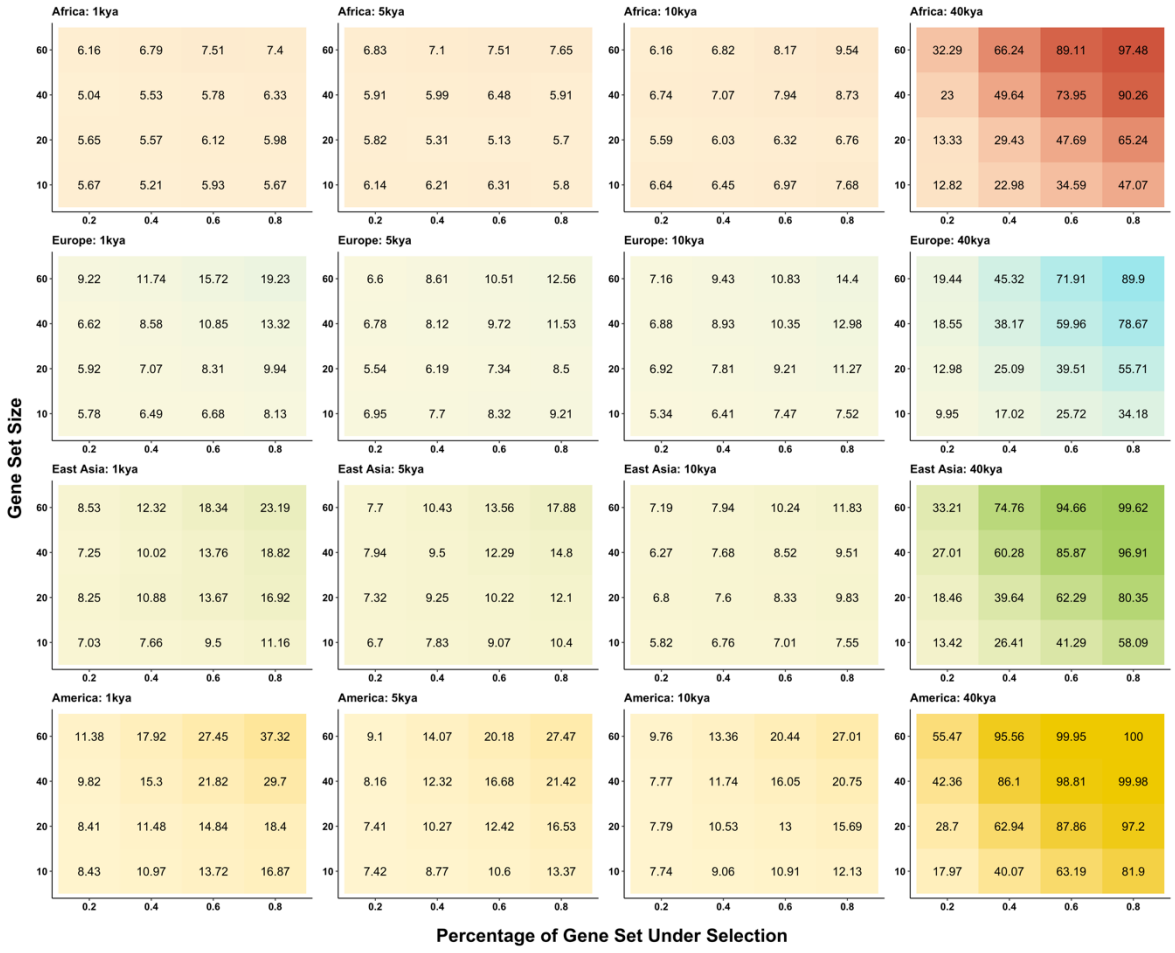


***Fig. S2.18**: The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating Relate values, partitioned by proportion of gene set under selection. Shown for selection acting on gene set sizes of 10, 20, 40 and 60, acting at 1kya, 5kya, 10kya, 40kya on four different populations.*

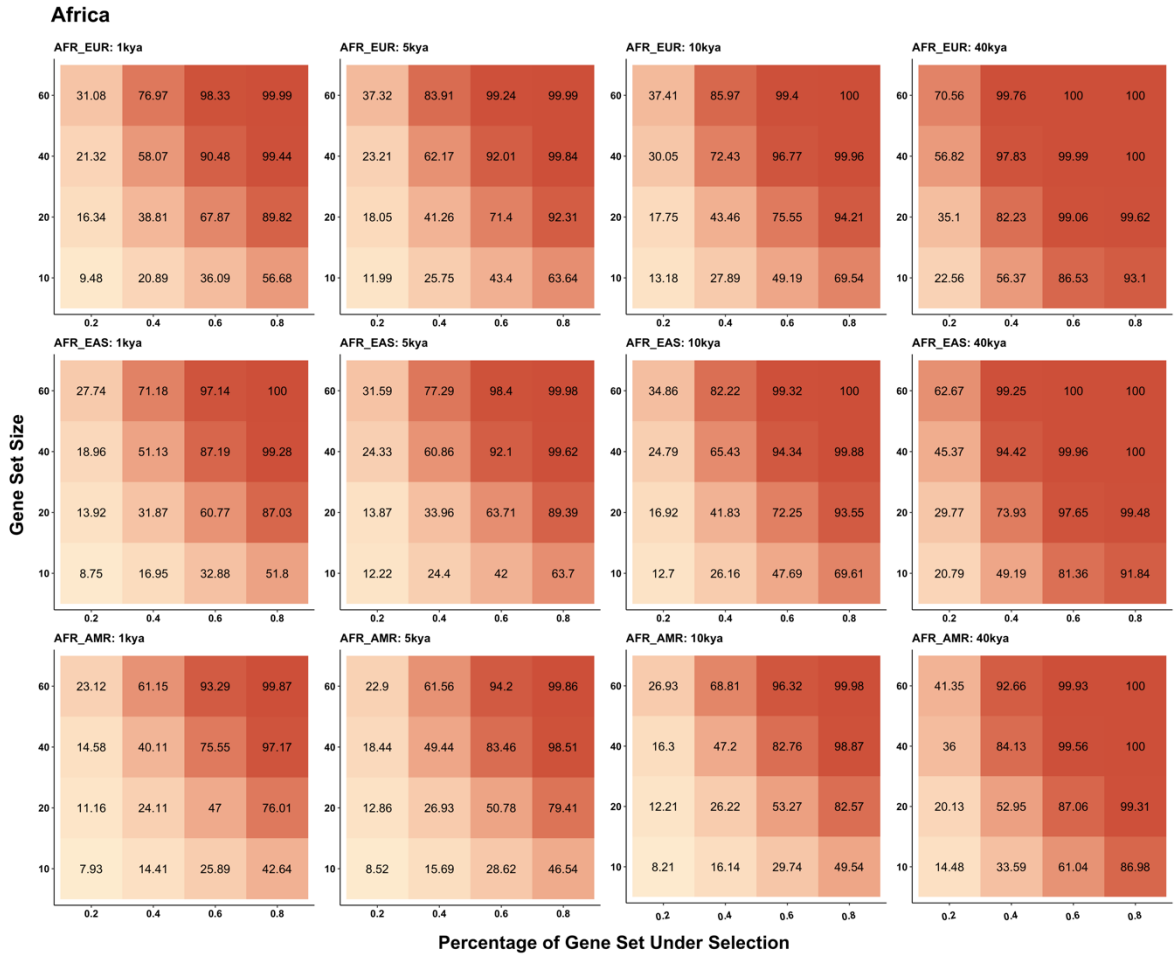


Fig. S2.19: The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating African F_{ST} values partitioned by proportion of gene set under selection. *Shown for selection acting on the African population at 1kya, 5kya, 10kya, 40kya, on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

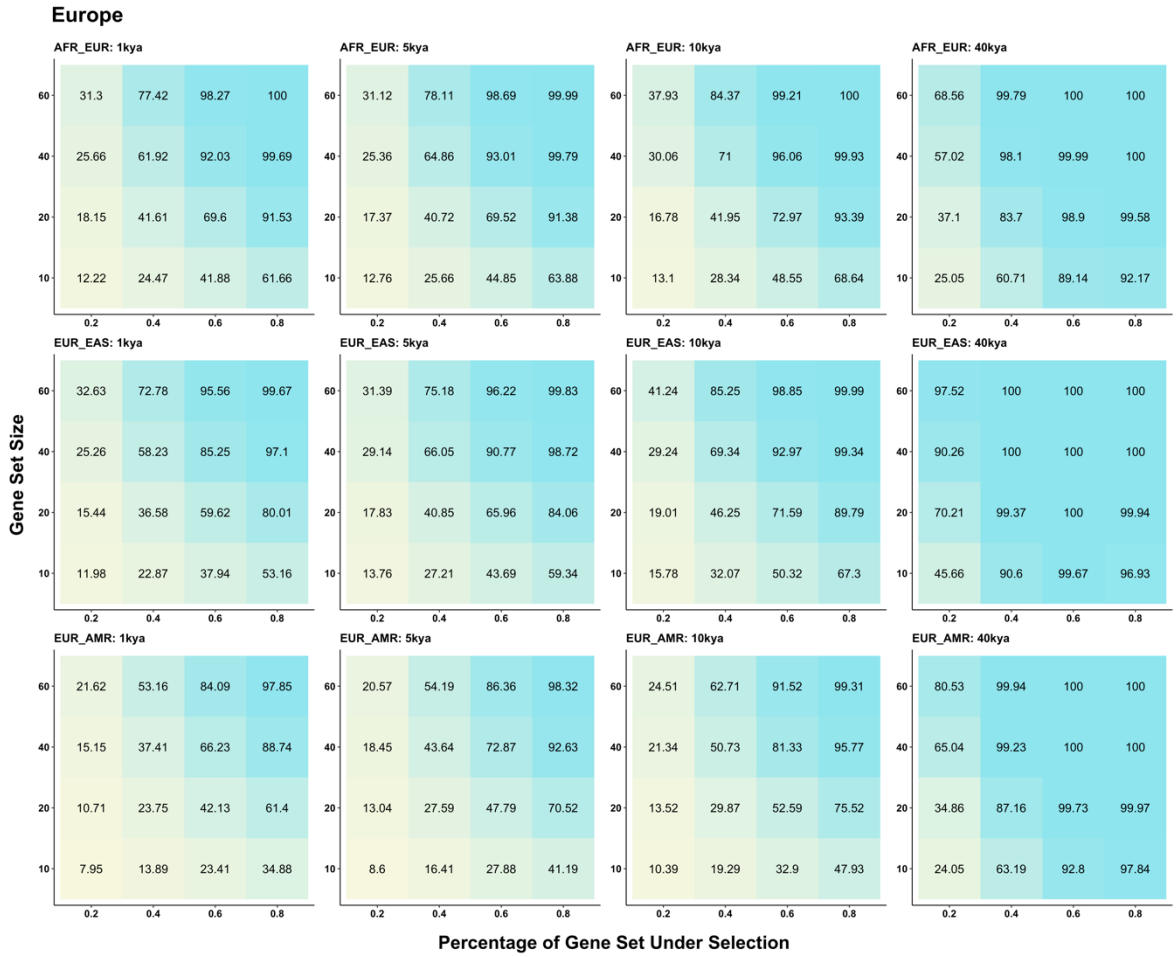


***Fig. S2.20**: **The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating European F_{ST} values partitioned by proportion of gene set under selection.** Shown for selection acting on the European population at 1kya, 5kya, 10kya, 40kya, on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

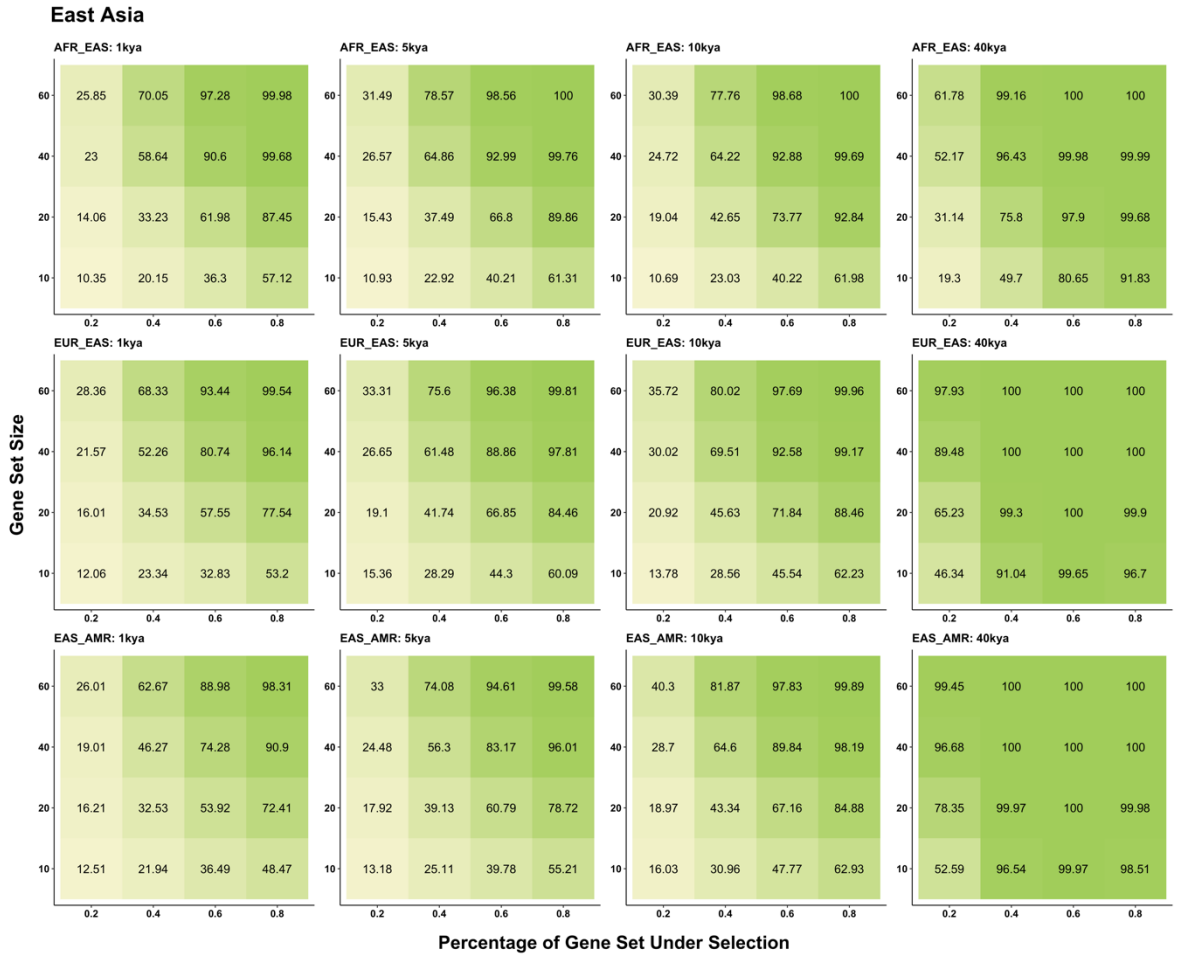


***Fig. S2.21**: **The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating East Asian F_{ST} values partitioned by proportion of gene set under selection.** Shown for selection acting on the East Asian population at 1kya, 5kya, 10kya, 40kya, on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

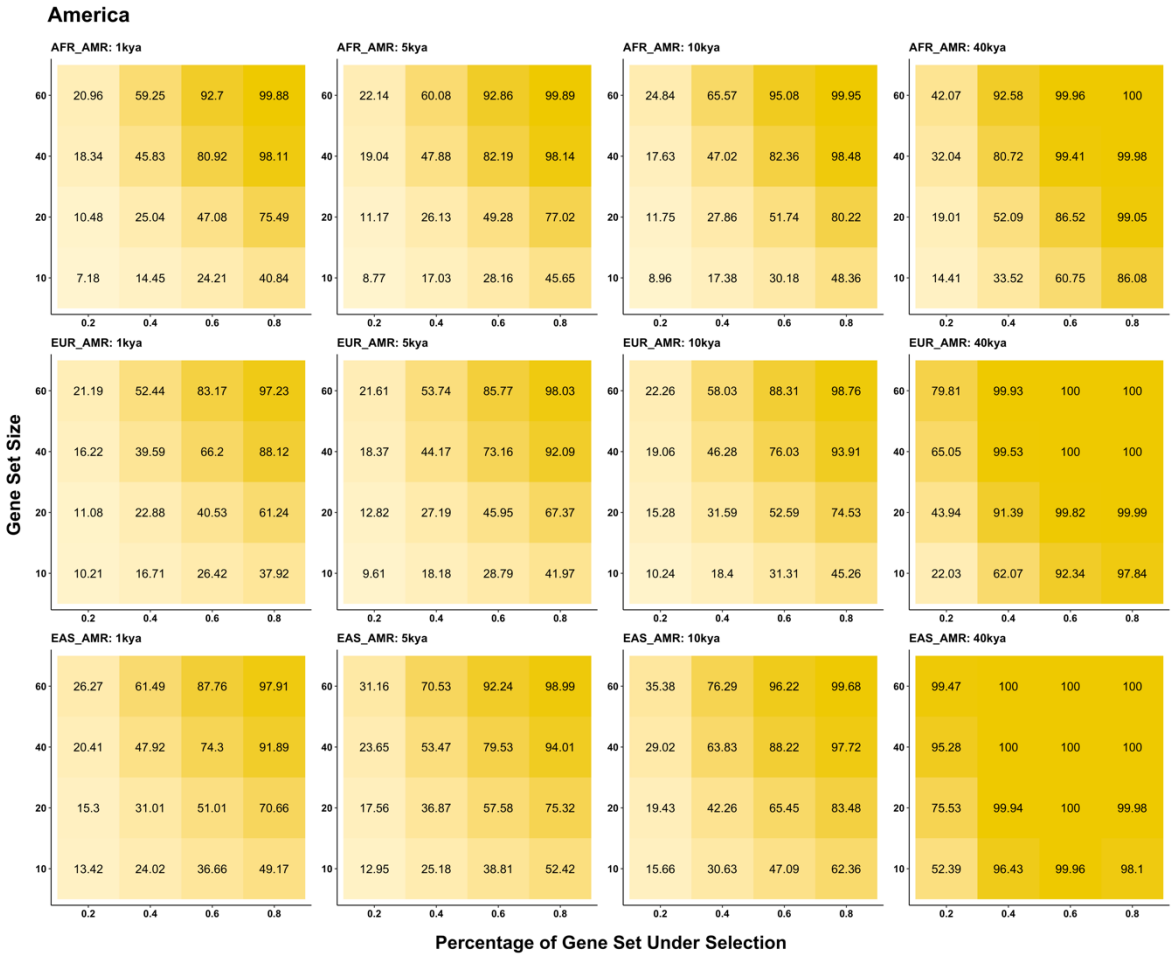


***Fig. S2.22**: **The percentage of gene sets identified as being under selection, according to the SUMSTAT method integrating American F_{ST} values partitioned by proportion of gene set under selection.** Shown for selection acting on the American population at 1kya, 5kya, 10kya, 40kya, on gene set sizes of 10, 20, 40 and 60. Given for three population comparisons, where AFR=Africa, EUR=Europe, EAS=East Asia, AMR=America.*

Chapter 3: Supplementary Material

Notes

***Note S3.1**: The code used to identify MA-gene sets with SUMSTAT summed values in the 5% tail of the background distribution. As generated from the SUMSTAT summed values generated from 1,000 neutral gene sets containing pMA-genes.*

```
import pandas as pd
import numpy as np
from scipy.stats import norm

pops=["BantuSouthAfrica_BantuKenya", "Biaka","Yoruba",
"Mandenka", "Mbuti", "San", "Bedouin", "Druze", "Mozabite",
"Palestinian", "Adygei", "Basque", "BergamoItalian_Tuscan",
"French", "Orcadian", "Russian", "Sardinian", "Balochi",
"Brahui", "Burusho", "Hazara", "Kalash", "Makrani", "Pathan",
"Sindhi", "Uygur", "Dai_Lahu", "Han", "Japanese",
"Oroqen_Hezhen_Daur", "Naxi_Yi", "NorthernHan_Tu",
"She_Miao_Tujia", "Xibo_Mongolian", "Yakut", "Maya", "Pima",
"Surui_Karitiana", "Bougainville",
"PapuanHighlands_PapuanSepik"]

newfile=[]
for x in pops:
  dist =
pd.read_csv("/home/ssd/jrees/significant/relate/sumstat/all/{}
_neutral_summed".format(x), header=None)
  mean = np.mean(dist[0])
  std = np.std(dist[0])


file=pd.read_csv("/home/ssd/jrees/significant/relate/sumstat/a
ll_micros_summed_pop", header=None, sep=" ")
  score = file.loc[file[0]=="{}".format(x)][1].values[0]
  prob = norm(mean, std).cdf(score)
  newfile.append(prob)

file['Probability'] = newfile
np.savetxt("/home/ssd/jrees/significant/relate/sumstat/all_mic
ros_summed_pop", file, fmt = '%s %f %f',header="Population Sum
Probability", comments="")
```

Tables

***Table S3.1*: All micronutrient-associated genes used in this study associated with the uptake, metabolism or regulation of 13 micronutrients.** *When genes are associated with multiple micronutrients, their most supported association given in "Micronutrient" with secondary or tertiary associations given in "Other Associations". Genes removed following the positive mask (Bergström et al., 2020) indicated in the "Removed During Pruning" column. Gene regions as taken from ensemble (Yates et al., 2020) and suggested from the literature ("Reference").*

Micronutrient	Gene Name	Gene Abbreviation	Chr.	Gene Start	Gene End	Other Associations	Evidence of selection	Removed During Pruning	Primary Ref
Selenium	Glutathione Peroxidase 1	*GPX1*	3	49357176	49358358		Yes		(White et al., 2015)
Selenium	Glutathione Peroxidase 2	*GPX2*	14	64939152	64942905				(White et al., 2015)
Selenium	Glutathione Peroxidase 3	*GPX3*	5	151020438	151028992		Yes		(White et al., 2015)
Selenium	Glutathione Peroxidase 4	*GPX4*	19	1103926	1106791				(White et al., 2015)
Selenium	Glutathione Peroxidase 6	*GPX6*	6	28503296	28528215				(White et al., 2015)
Selenium	Iodothyronine Deiodinase 1	*DIO1*	1	53891239	53911086	iodine			(White et al., 2015)
Selenium	Iodothyronine Deiodinase 2	*DIO2*	14	80197526	80387757	iodine	Yes		(White et al., 2015)
Selenium	Iodothyronine Deiodinase 3	*DIO3*	14	101561351	101563452	iodine			(White et al., 2015)
Selenium	Selenoprotein F	*SELENOF*	1	86862445	86914424				(White et al., 2015)
Selenium	Selenoprotein H	*SELENOH*	11	57741250	57743554				(White et al., 2015)
Selenium	Selenoprotein I	*SELENOI*	2	26308547	26395891				(White et al., 2015)
Selenium	Selenoprotein K	*SELENOK*	3	53884417	53891962				(White et al., 2015)
Selenium	Selenoprotein M	*SELENOM*	22	31104772	31120069				(White et al., 2015)
Selenium	Selenoprotein N	*SELENON*	1	25800176	25818221				(White et al., 2015)
Selenium	Selenoprotein O	*SELENOO*	22	50200979	50217616				(White et al., 2015)
Selenium	Selenoprotein T	*SELENOT*	3	150602875	150630445				(White et al., 2015)
Selenium	Selenoprotein V	*SELENOV*	19	39515113	39520686				(White et al., 2015)
Selenium	Selenoprotein W	*SELENOW*	19	47778585	47784686				(White et al., 2015)
Selenium	Methionine Sulfoxide Reductase B1	*MSRB1*	16	1938210	1943326				(White et al., 2015)
Selenium	Thioredoxin Reductase 1	*TXNRD1*	12	104215779	104350307				(White et al., 2015)
Selenium	Thioredoxin Reductase 2	*TXNRD2*	22	19875517	19941820				(White et al., 2015)
Selenium	Thioredoxin Reductase 3	*TXNRD3*	3	126607059	126655124				(White et al., 2015)
Selenium	Glutathione Peroxidase 5	*GPX5*	6	28525881	28534955			Yes	(White et al., 2015)
Selenium	Glutathione Peroxidase 7	*GPX7*	1	52602371	52609051				(White et al., 2015)
Selenium	Glutathione Peroxidase 8	*GPX8*	5	55160167	55167297				(White et al., 2015)
Selenium	Selenoprotein P	*SELENOP*	5	42799880	42887392				(White et al., 2015)
Selenium	LDL Receptor Related Protein 8	*LRP8*	1	53242364	53328469				(White et al., 2015)
Selenium	LDL Receptor Related Protein 2	*LRP2*	2	169127109	169362534				(White et al., 2015)
Selenium	Selenocysteine Lyase	*SCLY*	2	238060924	238099413				(White et al., 2015)
Selenium	Selenium Binding Protein 1	*SELENBP1*	1	151364304	151372707	copper	Yes		(White et al., 2015)
Selenium	Phosphoseryl-TRNA Kinase	*PSTK*	10	122954381	122997513				(White et al., 2015)
Selenium	O-phosphoseryl-tRNA(Sec) Selenium Transferase	*SEPSECS*	4	25120014	25160449				(White et al., 2015)

Selenium	Seryl-TRNA Synthetase 2	*SARS2*	19	38915266	38930763				(White et al., 2015)
Selenium	TRNA-SeC (Anticodon TCA) 1-1	*TRU-TCA1-1*	19	45478602	45478687				(White et al., 2015)
Selenium	TRNA-SeC (Anticodon TCA) 2-1	*TRU-TCA2-1*	22	44150657	44150742				(White et al., 2015)
Selenium	TRNA-SeC (Anticodon TCA) 3-1	*TRU-TCA3-1*	17	40117300	40117373				(White et al., 2015)
Selenium	CUGBP Elav-Like Family Member 1	*CELF1*	11	47465933	47565569		Yes		(White et al., 2015)
Selenium	Eurokaryotic Elongation Factor, Selenocysteine-TRNA Specific	*EEFSEC*	3	128153481	128408646				(White et al., 2015)
Selenium	Eukaryotic Translation Initiation factor 4A3	*EIF4A3*	17	80134369	80147151				(White et al., 2015)
Selenium	ELAV like RNA Binding Protein 1	*ELAVL1*	19	7958573	8005659				(White et al., 2015)
Selenium	Ribosomal Protein L30	*RPL30*	8	98024851	98046469				(White et al., 2015)
Selenium	SECIS Binding Protein 2	*SECISBP2*	9	89318500	89359663	iodine			(White et al., 2015)
Selenium	Selenophosphate synthetase 1	*SEPHS1*	10	13317428	13348298				(White et al., 2015)
Selenium	TRNA Selenocystein 1 Associated Protein 1	*TRNAU1AP*	1	28553085	28578545				(White et al., 2015)
Selenium	Exportin 1	*XPO1*	2	61477849	61538626				(White et al., 2015)
Selenium	A-Kinase Anchoring Protein 6	*AKAP6*	14	32329298	32837684				(Engelken et al., 2016)
Selenium	Fatty Acid Binding Protein 1	*FABP1*	2	88122982	88128062				(Engelken et al., 2016)
Selenium	Calcium-activated Potassium Channel Subfamily M Alpha-1	*KCNMA1*	10	76869601	77638369				(Engelken et al., 2016)
Selenium	Protein kinase CGMP-Dependent 1	*PRKG1*	10	50990888	52298423				(Engelken et al., 2016)
Selenium	Selenoprotein S	*SELENOS*	15	101270817	101277500		Yes		(Engelken et al., 2016)
Selenium	Selenoprotein Synthetase 2	*SEPHS2*	16	30443631	30445874		Yes		(Engelken et al., 2016)
Selenium	Sarcoglycan Delta	*SGCD*	5	155870344	156767788				(Engelken et al., 2016)
Selenium	Thioredoxin	*TXN*	9	110243810	110256507				(Engelken et al., 2016)
Selenium	Aldo-Keto Reductase Family 7 Like	*AKR7L*	1	19265982	19274194				(Wishart et al., 2007)
Selenium	Cystathionine Beta-Synthase	*CBS*	21	43053191	43076943			Yes	(Dib et al., 2019)
Selenium	Arylsulfatase B	*ARSB*	5	78777209	78986087				(Dib et al., 2019)
Selenium	LHFPL Tetraspan Subfamily Member 2	*LHFPL2*	5	78485215	78770021				(Dib et al., 2019)
Selenium	Dimethylglycine Dehydrogenase	*DMGDH*	5	78997564	79236038				(Dib et al., 2019)
Selenium	Betaine-Homocysteine S-Methyltransferase 2	*BHMT2*	5	79069767	79090069				(Dib et al., 2019)
Selenium	Betaine-Homocysteine S-Methyltransferase 2	*BHMT*	5	79111809	79132288				(Dib et al., 2019)
Selenium	Junction Mediating And Regulatory Protein, P53 Cofactor	*JMY*	5	79236131	79327211				(Dib et al., 2019)
Copper	Antioxidant 1 Copper Chaperone	*ATOX1*	5	151742316	151772532				(Engelken et al., 2016)
Copper	ATPase Copper Transporting Alpha	*ATP7A*	X	77910656	78050395				(Engelken et al., 2016)
Copper	ATPase Copper Transporting Beta	*ATP7B*	13	51930436	52012125				(Engelken et al., 2016)
Copper	Copper Metabolism Domain Containing 1	*COMMD1*	2	61888724	62147247				(Engelken et al., 2016)
Copper	X-linked Inhibitor Of Apoptosis	*XIAP*	X	123859724	123913979				(Wishart et al., 2007)
Copper	Solute Carrier Family 31 Member 1	*SLC31A1*	9	113221544	113264492				(Engelken et al., 2016)
Copper	Solute Carrier Family 31 Member 2	*SLC31A2*	9	113150976	113164140				(Engelken et al., 2016)
Copper	Superoxide Dismutase 1	*SOD1*	21	31659666	31668931				(Engelken et al., 2016)
Copper	Coiled-Coil Domain Containing 27	*CCDC27*	1	3746460	3771645				(Dib et al., 2019)
Iron	3-Hydroxybutyrate Dehydrogenase 2	*BDH2*	4	103077592	103099870				(Engelken et al., 2016)
Iron	Cytochrome D Reductase 1	*CYBRD1*	2	171522247	171558129				(Engelken et al., 2016)
Iron	Endothelial PAS Domain Protein 1	*EPAS1*	2	46293667	46386697		Yes		(Engelken et al., 2016)
Iron	Ferrochelatase	*FECH*	18	57544377	57586702				(Engelken et al., 2016)

Iron	Ferritin Heavy Chain 1	*FTH1*	11	61959718	61967634				(Engelken et al., 2016)
Iron	Ferritin Light Chain	*FTL*	19	48965309	48966879				(Engelken et al., 2016)
Iron	Hepcidin Antimicrobial Peptide	*HAMP*	19	35280716	35285143				(Engelken et al., 2016)
Iron	Hephaestin	*HEPH*	X	66162549	66268867				(Engelken et al., 2016)
Iron	Homeostatic Iron Regulator	*HFE*	6	26087281	26098343		Yes		(Engelken et al., 2016)
Iron	Hemojuveline BMP Co-Receptor	*HJV*	1	146017468	146036746				(Engelken et al., 2016)
Iron	Hypoxia Inducible Factor 1 Subunit Alpha	*HIF1A*	14	61695513	61748259				(Engelken et al., 2016)
Iron	Lactotransferrin	*LTF*	3	46435645	46485234				(Engelken et al., 2016)
Iron	Ras Homolog Family Member A	*RHOA*	3	49359145	49412998				(Engelken et al., 2016)
Iron	Solute Carrier Family 17 Member 1	*SLC17A1*	6	25782915	25832052				(Engelken et al., 2016)
Iron	Solute Carrier Family 40 Member 1	*SLC40A1*	2	189560590	189583758				(Engelken et al., 2016)
Iron	STEAP3 Metalloreductase	*STEAP3*	2	119223831	119265652				(Engelken et al., 2016)
Iron	Transferrin	*TF*	3	133746040	133796641				(Engelken et al., 2016)
Iron	Transferrin Receptor 2	*TFR2*	7	100620416	100642779				(Engelken et al., 2016)
Iron	Transferrin Receptor	*TFRC*	3	196027183	196082096				(Engelken et al., 2016)
Iron	Transmembrane Serine Protease 6	*TMPRSS6*	22	37065436	37109713				(Engelken et al., 2016)
Iron	Iron-Sulfur Cluster Assembly Enzyme	*ISCU*	12	108562582	108569384				(Engelken et al., 2016)
Iron	Lipcalin 2	*LCN2*	9	128149071	128153453				(Engelken et al., 2016)
Iron	Ferritin Mitochondrial	*FTMT*	5	121851882	121852833				(Wishart et al., 2007)
Iron	Aconitase 1	*ACO1*	9	32384603	32454769				(Muckenthaler et al., 2008)
Iron	Aconitase 2	*ACO2*	22	41469117	41528989				(Muckenthaler et al., 2008)
Iron	5'-Aminolevulinate Synthase 2	*ALAS2*	X	55009055	55030977				(Muckenthaler et al., 2008)
Iron	Solute Carrier Family 46 Member 1	*SLC46A1*	17	28394642	28407197				(Muckenthaler et al., 2008)
Iron	Solute Carrier Family 11 Member 1	*SLC11A1*	2	218382029	218396894	zinc			(Fishilevich et al., 2017)
Iron	Solute Carrier Family 48 Member 1	*SLC48A1*	12	47753916	47782751				(Fishilevich et al., 2017)
Iron	Solute Carrier Family 11 Member 2	*SLC11A2*	12	5097940151028566		magnesium			(Muckenthaler et al., 2008)
Iron	HBS1 Like Translational GTPase	*HBS1L*	6	134960378	135103056				(Dib et al., 2019)
Iron	MYB Proto-Oncogene	*MYB*	6	135181308	135219173				(Dib et al., 2019)
Iron	Phosphatidylinositol-4,5-Bisphosphate 3-Kinase Catalytic Subunit Gamma	*PIK3CG*	7	106865278	106908980				(Dib et al., 2019)
Iron	Cilia and Flagella Associated Protein 251	*CFAP251*	12	121918592	122003927				(Dib et al., 2019)
Iron	RHO Guanine Nucleotide Exchange Factor 3	*ARHGEF3*	3	56727418	57079329				(Dib et al., 2019)
Iron	TAO Kinase 1	*TAOK1*	17	29390464	29551904				(Dib et al., 2019)
Iron	Ceruloplasmin	*CP*	3	149162410	149221829	copper			(Fishilevich et al., 2017)
Iron	Pantothenate Kinase 2	*PANK2*	20	3888839	3929882				(Fishilevich et al., 2017)
Iron	Phospholipase A2 Group 6	*PLA2G6*	22	38111495	38214778				(Fishilevich et al., 2017)
Iron	Chromosome 19 Open Reading Frame 12	*C19ORF12*	19	29698886	29715789				(Fishilevich et al., 2017)
Iron	Fatty Acid 2-Hydroxylase	*FA2H*	16	74712955	74774831				(Fishilevich et al., 2017)
Iron	WD Repeat Domain 45	*WDR45*	X	49074433	49101170				(Fishilevich et al., 2017)
Iron	ATPase Cation Transporting 13A2	*ATP13A2*	1	16985958	17011928	manganese			(Fishilevich et al., 2017)
Magnesium	Solute Carrier Family 41 Member 1	*SLC41A1*	1	205789094	205813748				(Engelken et al., 2016)

Magnesium	Transient Receptor Potential Cation Channel Subfamily M Member 6	*TRPM6*	9	74722495	74888094				(Houillier, 2014)
Magnesium	Claudin 16	*CLDN16*	3	190322541	190412143				(Houillier, 2014)
Magnesium	Claudin 19	*CLDN19*	1	42733093	42740254				(Houillier, 2014)
Magnesium	Potassium Voltage-Gated Channel Subfamily A Member 1	*KCNA1*	12	4909905	4918256				(Houillier, 2014)
Magnesium	Cyclin and CBS Domain Divalent Metal Cation Transport Mediator 2	*CNNM2*	10	102918294	103090222				(Houillier, 2014)
Magnesium	FXYD Domain Containing Ion Transport Regulator 2	*FXYD2*	11	117800844	117828698				(Houillier, 2014)
Magnesium	Mitochondrial E3 Ubiquitin Protein Ligase 1	*MUL1*	1	20499448	20508151				(Houillier, 2014)
Magnesium	Doublecortin Domain Containing 1	*DCDC1*	11	30830369	31369810				(Houillier, 2014)
Magnesium	Shroom Family Member 3	*SHROOM3*	4	76435229	76783253				(Houillier, 2014)
Magnesium	MDS1 and EVI1 Complex Locus	*MECOM*	3	169083499	169663775				(Houillier, 2014)
Magnesium	Fibroblast Growth Factor Receptor 2	*FGFR2*	10	121478332	121598458				(Houillier, 2014)
Magnesium	3′-Phosphoadenosine 5′-Phosphosulfate Synthase 2	*PAPSS2*	10	87659613	87747705				(Houillier, 2014)
Magnesium	ADP Ribosylation Factor like GTPase 15	*ARL15*	5	53883942	54310582				(Houillier, 2014)
Magnesium	Epidermal Growth Factor	*EGF*	4	109912883	110013766				(Fishilevich et al., 2017)
Zinc	G Protein-Coupled Receptor 39	*GPR39*	2	132416805	132646582				(Engelken et al., 2016)
Zinc	Interleukin 6	*IL6*	7	22725884	22732002				(Engelken et al., 2016)
Zinc	Interleukin 6 Receptor	*IL6R*	1	154405193	154469450				(Engelken et al., 2016)
Zinc	Metallothionein 1A	*MT1A*	16	56638666	56640087				(Engelken et al., 2016)
Zinc	Metallothionein 1E	*MT1E*	16	56625475	56627112				(Engelken et al., 2016)
Zinc	Metallothionein 1F	*MT1F*	16	56657731	56660698				(Engelken et al., 2016)
Zinc	Metallothionein 1G	*MT1G*	16	56666730	56668065				(Engelken et al., 2016)
Zinc	Metallothionein 1H	*MT1H*	16	56669814	56671129				(Engelken et al., 2016)
Zinc	Metallothionein 2A	*MT2A*	16	56608584	56609497				(Engelken et al., 2016)
Zinc	Metallothionein 4	*MT4*	16	56565073	56568957				(Engelken et al., 2016)
Zinc	Metal Response Element Binding Transcription Factor 1	*MTF1*	1	37809574	37859592				(Engelken et al., 2016)
Zinc	Metal Response Element Binding Transcription Factor 2	*MTF2*	1	93079235	93139079				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 1	*SLC30A1*	1	211571568	211579161				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 2	*SLC30A2*	1	26037252	26046118				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 3	*SLC30A3*	2	27253684	27275817				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 4	*SLC30A4*	15	45479606	45522755				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 5	*SLC30A5*	5	69093949	69131069				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 6	*SLC30A6*	2	32165841	32224379				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 7	*SLC30A7*	1	100896076	100981757				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 8	*SLC30A8*	8	116950273	117176714				(Engelken et al., 2016)
Zinc	Solute Carrier Family 30 Member 9	*SLC30A9*	4	41990502	42090461		Yes		(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 1	*SLC39A1*	1	153959099	153968184				(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 10	*SLC39A10*	2	195575977	195737702				(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 11	*SLC39A11*	17	72645949	73092712				(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 12	*SLC39A12*	10	17951839	18043292				(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 13	*SLC39A13*	11	47407132	47416496				(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 2	*SLC39A2*	14	20999255	21001871				(Engelken et al., 2016)

Zinc	Solute Carrier Family 39 Member 3	*SLC39A3*	19	2732204	2740028				(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 4	*SLC39A4*	8	144409742	144416844		Yes		(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 5	*SLC39A5*	12	56230049	56237846		Yes		(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 6	*SLC39A6*	18	36108531	36129385				(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 7	*SLC39A7*	6	33200445	33204439		Yes	Yes	(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 8	*SLC39A8*	4	102251041	102431258	magnesium, manganese	Yes		(Engelken et al., 2016)
Zinc	Solute Carrier Family 39 Member 9	*SLC39A9*	14	69398015	69462390				(Engelken et al., 2016)
Zinc	Signal Transducer and Activator of Transcription 3	*STAT3*	17	42313324	42388568				(Engelken et al., 2016)
Zinc	Carbonic Anhydrase 1	*CA1*	8	85327608	85379014				(Dib et al., 2019)
Zinc	Carbonic Anhydrase 2	*CA2*	8	85463968	85481493				(Dib et al., 2019)
Zinc	Carbonic Anhydrase 3	*CA3*	8	85373436	85449040				(Dib et al., 2019)
Zinc	Carbonic Anhydrase 13	*CA13*	8	85220587	85284073				(Dib et al., 2019)
Zinc	Secretory Carrier Membrane Protein 5	*SCAMP5*	15	74957219	75021495				(Dib et al., 2019)
Zinc	KLF Transcription Factor 8	*KLF8*	X	56232356	56291531				(Dib et al., 2019)
Zinc	Zinc Finger X-Linked Duplicated A	*ZXDA*	X	57906708	57910820				(Dib et al., 2019)
Zinc	Zinc Finger X-Linked Duplicated B	*ZXDB*	X	57591652	57597545				(Dib et al., 2019)
Sodium	Sodium Channel Epithelial 1 Subunit Alpha	*SCNN1A*	12	6346843	6377730	potassium			(Engelken et al., 2016)
Sodium	Sodium Channel Epithelial 1 Subunit Beta	*SCNN1B*	16	23278231	23381294	potassium			(Rossier et al., 2002)
Sodium	Sodium Channel Epithelial 1 Subunit Delta	*SCNN1D*	1	1280436	1292029	potassium			(Rossier et al., 2002)
Sodium	Sodium Channel Epithelial 1 Subunit Gamma	*SCNN1G*	16	23182745	23216883	potassium			(Rossier et al., 2002)
Sodium	Nuclear Receptor Subfamily 3 Group C Member 2	*NR3C2*	4	148078762	148444698				(Rossier et al., 2002)
Sodium	Angiotensinogen	*AGT*	1	230702523	230714122				(Rossier et al., 2002)
Sodium	FXYD Domain Containing Ion Transport Regulator 4	*FXYD4*	10	43371636	43376335				(Rossier et al., 2002)
Sodium	FXYD Domain Containing Ion Transport Regulator 3	*FXYD3*	19	35115879	35124324				(Rossier et al., 2002)
Sodium	FXYD Domain Containing Ion Transport Regulator 1	*FXYD1*	19	35138808	35143109				(Rossier et al., 2002)
Sodium	FXYD Domain Containing Ion Transport Regulator 5	*FXYD5*	19	35154730	35169881				(Rossier et al., 2002)
Sodium	FXYD Domain Containing Ion Transport Regulator 7	*FXYD7*	19	35143250	35154302				(Rossier et al., 2002)
Sodium	Sodium Voltage-Gated Channel Beta Subunit 3	*SCN3B*	11	123629187	123655244				(Rossier et al., 2002)
Sodium	NEDD4 E3 Ubiquitin Protein Ligase	*NEDD4*	15	55826922	55993746				(Rossier et al., 2002)
Sodium	Serum/Glucocorticoid Regulated Kinase 1	*SGK1*	6	134169246	134318112				(Rossier et al., 2002)
Sodium	Serine/Threonine Kinase 39	*STK39*	2	167954020	168247595				(Freitas, 2018)
Sodium	G Protein-Coupled Receptor Kinase 4	*GRK4*	4	2963571	3040760				(Freitas, 2018)
Sodium	Solute Carrier Family 4 Member 5	*SLC4A5*	2	74216242	74343414				(Freitas, 2018)
Calcium	Transient Receptor Potential Cation Channel Subfamily M Member 2	*TRPM2*	21	44350163	44443081				(Engelken et al., 2016)
Calcium	Transient Receptor Potential Cation Channel Subfamily V Member 5	*TRPV5*	7	142908101	142933746			Yes	(Kovacs et al., 2013)
Calcium	Transient Receptor Potential Cation Channel Subfamily V Member 6	*TRPV6*	7	142871208	142885745		Yes	Yes	(Hughes et al., 2008)
Calcium	Calcium Sensing Receptor	*CASR*	3	122183668	122291629	magnesium, phosphorus			(Houillier, 2014)
Calcium	B-Box and SPRY Domain Containing	*BSPRY*	9	113349541	113371233				(Khanal & Nemere, 2008)
Calcium	Regulator of G Protein Signalling 2	*RGS2*	1	192809039	192812275				(Khanal & Nemere, 2008)
Calcium	Solute Carrier Family 8 Member 1	*SLC8A1*	2	40097270	40611053				(Khanal & Nemere, 2008)

Calcium	Solute Carrier Family 8 Member 2	*SLC8A2*	19	47428017	47471893				(Khanal & Nemere, 2008)
Calcium	Solute Carrier Family 8 Member 3	*SLC8A3*	14	70044215	70189070				(Khanal & Nemere, 2008)
Calcium	ATPase Plasma Membrane Ca2+ Transporting 1	*ATP2B1*	12	89588049	89709300				(Freitas, 2018)
Calcium	ATPase Plasma Membrane Ca2+ Transporting 2	*ATP2B2*	3	10324023	10708007				(Khanal & Nemere, 2008)
Calcium	ATPase Plasma Membrane Ca2+ Transporting 3	*ATP2B3*	X	153517676	153582939				(Khanal & Nemere, 2008)
Calcium	ATPase Plasma Membrane Ca2+ Transporting 4	*ATP2B4*	1	203626561	203744081				(Khanal & Nemere, 2008)
Calcium	Parathyroid Hormone	*PTH*	11	13492054	13496181	potassium			(Khanal & Nemere, 2008)
Calcium	Cytochrome P450 Family 24 Subfamily A Member 1	*CYP24A1*	20	54153446	54173986				(Dib et al., 2019)
Calcium	GATA Binding Protein 3	*GATA3*	10	8045378	8075198				(Dib et al., 2019)
Calcium	Diacylglycerol Kinase Delta	*DGKD*	2	233354507	233472104				(Dib et al., 2019)
Calcium	Von Willebrand Factor A Domain Containing 8	*VWA8*	13	41566835	41961120				(Dib et al., 2019)
Calcium	Glucokinase Regulator	*GCKR*	2	27496839	27523684				(Dib et al., 2019)
Iodine	Thyroid Hormone Receptor Interactor 4	*TRIP4*	15	64387748	64455303		Yes		(Herráez et al., 2009)
Iodine	Iodotyrosine Deiodinase	*IYD*	6	150368892	150405969		Yes		(Herráez et al., 2009)
Iodine	Solute Carrier Family 5 Member 5	*SLC5A5*	19	17871945	17895174	sodium			(Engelken et al., 2016)
Iodine	Solute Carrier Family 16 Member 10	*SLC16A10*	6	111087503	111231194				(The UniProt Consortium, 2023)
Iodine	Thyroid Hormone Receptor Alpha	*THRA*	17	40058290	40093867				(The UniProt Consortium, 2023)
Iodine	Thyroid Hormone Receptor Beta	*THRB*	3	24117153	24495756				(The UniProt Consortium, 2023)
Iodine	Solute Carrier Family 16 Member 2	*SLC16A2*	X	74421493	74533917				(The UniProt Consortium, 2023)
Iodine	Thyroid Stimulating Hormone Receptor	*TSHR*	14	80954989	81146302				(The UniProt Consortium, 2023)
Iodine	Solute Carrier Organic Anion Transporter Family Member 1C1	*SLCO1C1*	12	20695355	20753386				(The UniProt Consortium, 2023)
Iodine	Thyroid Peroxidase	*TPO*	2	1374066	1543711				(Wishart et al., 2007)
Iodine	Transthyretin	*TTR*	18	31557010	31599021				(Wishart et al., 2007)
Iodine	Serpin Family A Member 7	*SERPINA7*	X	106032435	106038727				(Wishart et al., 2007)
Iodine	Solute Carrier Family 3 Member 2	*SLC3A2*	11	62856102	62888875				(Wishart et al., 2007)
Iodine	Sulfotransferase Family 6B Member 1	*SULT6B1*	2	37167820	37196598				(Wishart et al., 2007)
Chloride	Chloride Voltage-Gated Channel 3	*CLCN3*	4	169612633	169723673				(Stauber & Jentsch, 2013)
Chloride	Chloride Voltage-Gated Channel 4	*CLCN4*	X	10156945	10237660				(Stauber & Jentsch, 2013)
Chloride	Chloride Voltage-Gated Channel 5	*CLCN5*	X	49922596	50099235				(Stauber & Jentsch, 2013)
Chloride	Chloride Voltage-Gated Channel 6	*CLCN6*	1	11806096	11848079				(Stauber & Jentsch, 2013)
Chloride	Chloride Voltage-Gated Channel 7	*CLCN7*	16	1444934	1475084				(Stauber & Jentsch, 2013)
Chloride	CF Transmembrane Conductance Regulator	*CFTR*	7	117287120	117715971				(Stauber & Jentsch, 2013)
Chloride	Aquaporin 6	*AQP6*	12	49967194	49977139				(Stauber & Jentsch, 2013)
Chloride	Anoctamin 3	*ANO3*	11	26188842	26663289				(Stauber & Jentsch, 2013)

Chloride	Anoctamin 4	*ANO4*	12	100717526	101128641				(Stauber & Jentsch, 2013)
Chloride	Anoctamin 5	*ANO5*	11	21799934	22283357				(Stauber & Jentsch, 2013)
Chloride	Anoctamin 6	*ANO6*	12	45215987	45440404				(Stauber & Jentsch, 2013)
Chloride	Anoctamin 7	*ANO7*	2	241188509	241225377				(Stauber & Jentsch, 2013)
Chloride	Bestrophin 1	*BEST1*	11	61950063	61965515				(Stauber & Jentsch, 2013)
Chloride	G Protein-Coupled Receptor 89A	*GPR89A*	1	145607988	145670650			Yes	(Stauber & Jentsch, 2013)
Chloride	Chloride Intracellular Channel 1	*CLIC1*	6	31730581	31739763			Yes	(Stauber & Jentsch, 2013)
Chloride	Chloride Intracellular Channel 2	*CLIC2*	X	155276211	155334657				(Stauber & Jentsch, 2013)
Chloride	Chloride Intracellular Channel 3	*CLIC3*	9	136994608	136996568				(Stauber & Jentsch, 2013)
Chloride	Chloride Intracellular Channel 4	*CLIC4*	1	24745382	24844321				(Stauber & Jentsch, 2013)
Chloride	Chloride Intracellular Channel 5	*CLIC5*	6	45880827	46080348				(Stauber & Jentsch, 2013)
Chloride	Chloride Intracellular Channel 6	*CLIC6*	21	34669389	34718227				(Stauber & Jentsch, 2013)
Chloride	Solute Carrier Family 17 Member 7	*SLC17A7*	19	49429401	49442360				(Stauber & Jentsch, 2013)
Chloride	Solute Carrier Family 12 Member 2	*SLC12A2*	5	128083766	128189677				(Wishart et al., 2007)
Chloride	Chloride Voltage-Gated Channel Kb	*CLCNKB*	1	16043736	16057308				(Jain et al., 2013)
Chloride	Barttin CLCNK Type Accessory Subunit Beta	*BSND*	1	54998933	55017172				(Jain et al., 2013)
Potassium	Potassium Inwardly Rectifying Channel Subfamily J Member 10	*KCNJ10*	1	159998651	160070160	calcium			(Jain et al., 2013)
Potassium	Cytochrome P450 Family 11 Subfamily B Member 1	*CYP11B1*	8	142872356	142879846				(Jain et al., 2013)
Potassium	Cytochrome P450 Family 11 Subfamily B Member 2	*CYP11B2*	8	142910559	142917843				(Jain et al., 2013)
Potassium	Hydroxysteroid 11-Beta Dehydrogenase 2	*HSD11B2*	16	67430652	67437553	sodium			(Jain et al., 2013)
Potassium	Solute Carrier Family 12 Member 1	*SLC12A1*	15	48178438	48304078	sodium, chloride			(Jain et al., 2013)
Potassium	Potassium Inwardly Rectifying Channel Subfamily J Member 1	*KCNJ1*	11	128836315	128867373				(Jain et al., 2013)
Potassium	Solute Carrier Family 12 Member 3	*SLC12A3*	16	56865207	56915850	calcium, magnesium			(Jain et al., 2013)
Phosphorus	Solute Carrier Family 34 Member 1	*SLC34A1*	5	177379235	177398848	calcium			(Chang & Anderson, 2017)
Phosphorus	Solute Carrier Family 34 Member 2	*SLC34A2*	4	25648011	25678748				(Chang & Anderson, 2017)
Phosphorus	Solute Carrier Family 34 Member 3	*SLC34A3*	9	137230757	137236555	calcium			(Chang & Anderson, 2017)
Phosphorus	Fibroblast Growth Factor 23	*FGF23*	12	4368227	4379712				(Chang & Anderson, 2017)
Phosphorus	Polypeptide N-Acetylgalactosaminyltransferase 3	*GALNT3*	2	165747588	165794659				(Chang & Anderson, 2017)
Phosphorus	Alkaline Phosphatase, Biomineralization Associated	*ALPL*	1	21509397	21578410				(Dib et al., 2019)
Phosphorus	NBPF Member 3	*NBPF3*	1	21440128	21485005				(Dib et al., 2019)
Phosphorus	Phosphodiesterase 7B	*PDE7B*	6	135851701	136195574				(Dib et al., 2019)
Phosphorus	LEM Domain Nuclear Envelope Protein 2	*LEMD2*	6	33771202	33789130				(Dib et al., 2019)
Phosphorus	Motilin	*MLN*	6	33794673	33804003				(Dib et al., 2019)
Phosphorus	Inositol 1,4,5-Triphosphate Receptor Type 3	*ITPR3*	6	33620365	33696574				(Dib et al., 2019)

Phosphorus	Mitochondrial Matrix Import Factor 23	*CCDC58*	3	122359591	122383231				(Dib et al., 2019)
Phosphorus	Fibroblast Growth Factor 6	*FGF6*	12	4428155	4445614				(Dib et al., 2019)
Phosphorus	RAD51 Associated Protein 1	*RAD51AP1*	12	4538798	4560048				(Dib et al., 2019)
Manganese	Superoxide Dismutase 2	*SOD2*	6	159669069	159745186				(Engelken et al., 2016)
Manganese	Solute Carrier Family 30 Member 10	*SLC30A10*	1	219685427	219958647	zinc, magnesium			(Dib et al., 2019)
Manganese	Cytochrome P450 12c1	*12C1*	3	130850595	131016712				(Horning et al., 2015)
Manganese	Solute Carrier Family 39 Member 14	*SLC39A14*	8	22367249	22434129	zinc			(Horning et al., 2015)
Molybdenum	Major Facilitator Superfamily Domain Containing 5	*MFSD5*	12	53251251	53254406				(Engelken et al., 2016)
Molybdenum	Molybdenum Cofactor Synthesis 1	*MOCS1*	6	39899578	39934551				(Reiss & Hahnewald, 2011)
Molybdenum	Molybdenum Cofactor Synthesis 2	*MOCS2*	5	53095679	53110063				(Reiss & Hahnewald, 2011)
Molybdenum	Molybdenum Cofactor Sulfurase	*MOCOS*	18	36187497	36272157				(Fishilevich et al., 2017)
Molybdenum	Gephyrin	*GPHN*	14	66507407	67181803				(Reiss & Hahnewald, 2011)

***Table S3.2**: All micronutrient-associated genes which are less than 10kbp ("Nature of Overlap" = "$\pm 10kbp$") or have overlapping gene regions as given by $ensemble$ ("Nature of Overlap" = "$ensemble$")*

Overlapping Genes				**Overlap (bp)**	**Nature of Overlap**
GPx1	(selenium)	RHOA	(iron)	787	$\pm 10kbp$
LHFPL2	(selenium)	ARSB	(selenium)	7188	$\pm 10kbp$
LEMD2	(phosphorus)	MLN	(phosphorus)	5543	$\pm 10kbp$
MT1F	(zinc)	MT1G	(zinc)	6032	$\pm 10kbp$
MT1G	(zinc)	MT1H	(zinc)	1749	$\pm 10kbp$
FXYD1	(sodium)	FXYD7	(sodium)	141	$\pm 10kbp$
FXYD7	(sodium)	FXYD5	(sodium)	428	$\pm 10kbp$
DMGDH	(selenium)	BHMT2	(selenium)	166271	$ensemble$
GPx5	(selenium)	GPx6	(selenium)	2334	$ensemble$
CA1	(zinc)	CA3	(zinc)	5578	$ensemble$
BEST1	(chloride)	FT1H	(zinc)	5797	$ensemble$

***Table S3.3**: **Details on the allele frequency distribution of all micronutrient-associated gene sets**. The number of SNPs over all genes in a given gene set (calculated from the Yoruba population), mean, median and standard deviation of the allele frequency distribution of all micronutrient-associated gene sets, the difference to that of the background (allele frequency distribution of chr1 of the Yoruba population; mean= 0.345, median= 0.227, standard deviation= 0.290), and the significance calculated when comparing these distributions (unpaired Wilcoxon test).*

	Number of SNPs	Mean	Difference to Background Mean	Median	Difference to Background Median	Standard Deviation	Difference to Background Standard Deviation	Significance
Selenium	17614	0.337	-0.008	0.227	0.000	0.282	-0.008	0.057
Copper	**1409**	**0.315**	**-0.030**	**0.182**	**-0.045**	**0.279**	**-0.011**	**0.0008543**
Iron	6476	0.343	-0.002	0.227	0.000	0.292	0.002	0.6131
Magnesium	**7862**	**0.356**	**0.011**	**0.227**	**0.000**	**0.293**	**0.003**	**0.03418**
Zinc	7755	0.356	0.011	0.227	0.000	0.297	0.007	0.1861
Sodium	**5016**	**0.327**	**-0.018**	**0.205**	**-0.023**	**0.284**	**-0.006**	**0.001019**
Calcium	6978	0.344	-0.001	0.205	-0.023	0.293	0.003	0.6302
Iodine	4035	0.351	0.006	0.250	0.023	0.290	0.000	0.2436
Chloride	8514	0.337	-0.008	0.227	0.000	0.281	-0.009	0.9225
Potassium	1682	0.342	-0.003	0.250	0.023	0.276	-0.014	0.1838
Phosphorus	2662	0.344	-0.001	0.205	-0.023	0.294	0.004	0.8334
Manganese	2152	0.348	0.003	0.205	-0.023	0.289	-0.001	0.08396
Molybdenum	**1390**	**0.427**	**0.082**	**0.273**	**0.045**	**0.305**	**0.015**	**2.20E-16**

***Table S3.4**: **Micronutrient genes enriched for SNP-density**. Enrichment given as over 95% quantile of the cumulative density function drawn from the distribution formed from generated neutral gene regions.*

Micronutrient	Gene	SNPs	Calculated CDF
Selenium	*SELENOO*	1083	0.99776
Iron	*EPAS1*	2829	0.97643
Zinc	*MT1A*	584	0.97112
Zinc	*MT1F*	631	0.98755
Sodium, Potassium	*SCNN1D*	807	0.95443
Calcium	*SLC8A1*	13155	0.98523
Chloride	*CLCN7*	1388	0.99038

***Table S3.5**: **The mean of the CDF position for each micronutrient gene set.** The significance of the difference given from a normal distribution centred at 0.5 (s.d.=0.25; wilcox-test). Drawn from the distribution formed from generated neutral gene regions.*

Micronutrient	Mean	Significance
Selenium	0.52796	0.09295
Copper	0.46716	0.3253
Iron	0.53543	0.6813
Magnesium	0.50585	0.9858
Zinc	0.48324	0.454
Sodium	0.59495	0.09706
Calcium	0.63266	0.01754
Iodine	0.50401	0.8116
Chloride	0.58836	0.1014
Potassium	0.54300	0.5248
Phosphorus	0.61845	0.03211
Manganese	0.41585	0.3144
Molybdenum	0.532746	0.6065

***Table S3.6**: **Populations used in this study.** Defined by (Bergström et al., 2020)*.

Metapopulation	Group name	Populations	Sample Size
Africa	Mbuti	Mbuti	13
	Biaka	Biaka	22
	San	San	6
	Bantu-speaking	Bantu(Kenya), Bantu(SouthAfrica)	19
	Yoruba	Yoruba	22
	Mandenka	Mandenka	22
Middle-East	Mozabite	Mozabite	27
	Palestinian	Palestinian	46
	Druze	Druze	42
	Bedouin	Bedouin	46
Europe	BergamoItalian-Tuscan	Bergamo_Italian, Tuscan	21
	Russian	Russian	25
	Adygei	Adygei	16
	Orcadian	Orcadian	15
	French	French	28
	Basque	Basque	23
	Sardinian	Sardinian	28
	Russian	Russian	25

East-Asia	Xibo-Mongolian	Mongolian, Xibo	18
	NorthernHan-Tu	NorthernHan, Tu	20
	Naxi-Yi	Naxi, Yi	18
	She-Miao-Tujia	She, Miao, Tujia	29
	Oroqen-Hezhen-Daur	Oroqen, Hezhen, Daur	27
	Dai-Lahu	Dai, Lahu	17
	Han	Han	33
	Japanese	Japanese	27
	Yakut	Yakut	25
Central-South Asia	Hazara	Hazara	19
	Uygur	Uygur	10
	Makrani	Makrani	25
	Sindhi	Sindhi	24
	Balochi	Balochi	24
	Brahui	Brahui	25
	Burusho	Barusho	25
	Kalash	Kalash	22
	Pathan	Pathan	24
Oceania	Papuan	Papuan (Sepik), Papuan(Highlands)	17
	Bougainville	Bougainville	11
Americas	Pima	Pima	13
	Maya	Maya	21
	Surui-Karitiana	Surui, Karitiana	20

***Table S3.7**: **Micronutrient-associated gene sets with significantly different summed selection values**. According to the gene set method SUMSTAT integrating Relate selection values. Partitioned by significance.*

Micronutrient	Population	Significance	
Phosphorus	Pima	0.000013	**< 0.0001**
Sodium	Adygei	0.000029	
Potassium	French	0.000322	**< 0.001**
Iodine	Maya	0.000325	
Sodium	Brahui	0.00115	**<0.01**
Potassium	BergamoItalian_Tuscan	0.002963	
Sodium	Bougainville	0.003455	
Potassium	Bougainville	0.003722	
Sodium	Russian	0.004935	
Sodium	Pathan	0.004951	
Sodium	San	0.0057	
Sodium	Orcadian	0.005823	
Sodium	French	0.006133	
Iodine	Mozabite	0.006333	
Calcium	Mozabite	0.007348	
Iodine	Russian	0.009037	
Sodium	NorthernHan_Tu	0.010827	**<0.05**
Sodium	BergamoItalian_Tuscan	0.01141	
Sodium	Basque	0.011611	
Potassium	NorthernHan_Tu	0.014566	
Sodium	Dai_Lahu	0.01592	

Potassium	Russian	0.017106
Potassium	Druze	0.018052
Calcium	Sardinian	0.018387
Calcium	Pima	0.018616
Sodium	Sindhi	0.021039
Potassium	Xibo_Mongolian	0.021087
Selenium	Xibo_Mongolian	0.02171
Magnesium	Surui_Karitiana	0.023716
Potassium	Mandenka	0.02482
Potassium	Sindhi	0.027062
Potassium	Palestinian	0.033025
Potassium	Mozabite	0.033461
Manganese	Naxi_Yi	0.034193
Zinc	Naxi_Yi	0.03529
Potassium	Sardinian	0.035508
Phosphorus	Yoruba	0.036495
Iodine	Orcadian	0.036859
Copper	Sardinian	0.037657
Calcium	Japanese	0.038349
Potassium	Kalash	0.038729
Phosphorus	PapuanHighlands_PapuanSepik	0.04091
Calcium	Maya	0.042026
Potassium	Pathan	0.044689
Calcium	Pathan	0.044822
Potassium	Yoruba	0.045052
Phosphorus	Pathan	0.046086
Calcium	Orcadian	0.047252

Table S3.8: Micronutrient-associated gene sets with significantly different summed selection values*. According to the gene set method SUMSTAT integrating F_{ST} selection values. Partitioned by significance.*

Micronutrient	Population	Significance	
Potassium	BantuSouthAfrica_BantuKenya	0.000043	**< 0.0001**
Sodium	Makrani	0.00048	**< 0.001**
Calcium	Mandenka	0.000912	
Calcium	Biaka	0.001264	**< 0.01**
Potassium	Orcadian	0.001556	
Potassium	Surui_Karitiana	0.001698	
Potassium	Russian	0.002343	
Zinc	Kalash	0.004891	
Potassium	Palestinian	0.005466	
Phosphorus	Mandenka	0.006715	
Sodium	Surui_Karitiana	0.0068	
Potassium	Mozabite	0.008791	
Potassium	French	0.0088	
Potassium	Kalash	0.009572	
Selenium	Xibo_Mongolian	0.00993	
Potassium	Pima	0.012051	**<0.05**
Sodium	French	0.013281	
Zinc	Uygur	0.016652	
Sodium	Orcadian	0.016658	
Sodium	Russian	0.017819	
Potassium	Basque	0.018501	
Potassium	Bedouin	0.01917	
Sodium	BergamoItalian_Tuscan	0.020302	
Potassium	Adygei	0.021671	
Iron	Mandenka	0.022027	
Potassium	Makrani	0.022717	
Potassium	Brahui	0.022756	
Sodium	Sindhi	0.024379	
Sodium	Basque	0.024715	
Selenium	Japanese	0.025141	
Potassium	Sardinian	0.026595	
Sodium	Brahui	0.026743	
Magnesium	Biaka	0.027088	

Sodium	Adygei	0.028808
Potassium	BergamoItalian_Tuscan	0.029975
Selenium	Pima	0.031199
Selenium	Surui_Karitiana	0.032171
Potassium	Sindhi	0.032643
Selenium	Han	0.033762
Selenium	She_Miao_Tujia	0.037054
Selenium	Oroqen_Hezhen_Daur	0.038023
Potassium	Balochi	0.038248
Potassium	Dai_Lahu	0.042516
Potassium	Burusho	0.042705
Sodium	Kalash	0.043516
Potassium	Pathan	0.046447
Sodium	Biaka	0.047979
Potassium	San	0.048742
Potassium	Maya	0.048774

Table S3.9: Micronutrient-associated gene sets with significantly different summed selection values*. According to the gene set method SUMSTAT integrating both Relate and F_{ST} selection values.*

Micronutrient	**Population**	*Relate* **Significance**	F_{ST} Significance
Selenium	Xibo-Mongolian	0.02171	0.00993
Sodium	Adygei	0.000029	0.028808
Sodium	Basque	0.011611	0.024715
Sodium	BergamoItalian-Tuscan	0.01141	0.020302
Sodium	French	0.006133	0.013281
Sodium	Orcadian	0.005823	0.016658
Sodium	Russian	0.004935	0.017819
Sodium	Brahui	0.00115	0.026743
Sodium	Sindhi	0.021039	0.024379
Potassium	Mozabite	0.033461	0.008791
Potassium	Palestinian	0.033025	0.005466
Potassium	BergamoItalian-Tuscan	0.002963	0.029975
Potassium	French	0.000322	0.0088
Potassium	Russian	0.017106	0.0088
Potassium	Sardinian	0.035508	0.026595
Potassium	Kalash	0.038729	0.009572
Potassium	Sindhi	0.027062	0.032643
Potassium	Pathan	0.044689	0.046447

Table S3.10*: Micronutrient-associated gene sets, as cut down to remove overlap, with significantly different summed selection values. According to the gene set method SUMSTAT integrating Relate selection values. Partitioned by significance.*

Micronutrient	**Population**	**Significance**	
Phosphorus	Pima	0.005012	**< 0.01**
Selenium	Xibo_Mongolian	0.02171	**< 0.05**

Table S3.11: Micronutrient-associated gene sets, as cut down to remove overlap, with significantly different summed selection values*. According to the gene set method SUMSTAT integrating F_{ST} selection values. Partitioned by significance.*

Micronutrient	Population	Significance	
Selenium	Xibo_Mongolian	0.00993	**< 0.01**
Iron	Mandenka	0.022027	
Selenium	Japanese	0.025141	
Selenium	Pima	0.031199	
Selenium	Surui_Karitiana	0.032171	
Selenium	Han	0.033762	
Selenium	She_Miao_Tujia	0.037054	
Selenium	Oroqen_Hezhen_Daur	0.038023	**< 0.05**

Table S3.12: The five MAGs for each population with the strongest evidence for selection, as indicated by Relate selection values*. When taking the only five MAGs would cut-off genes with the same significance value, more genes are given.*

Population	Gene	Micronutrient	Significance
San	*PRKG1*	selenium	0.0011774
	AKAP6	selenium	0.0011774
	SGCD	selenium	0.0011774
	SELENOP	selenium	0.0011774
	ATP7B	copper	0.0011774
	TSHR	iodine	0.0011774
	TRPM6	magnesium	0.0011774
	TXNRD3	selenium	0.0011774
	ANO4	chloride	0.0011774
	FECH	Iron	0.0011774
	SLC8A3	calcium	0.0011774
	STK39	sodium	0.0011774
	PSTK	selenium	0.0011774
Bantu-speaking	*SHROOM3*	magnesium	7.7e-6
	SLC39A11	zinc	3.3e-5
	GALNT3	phosphorus	5.52e-5
	LRP8	selenium	8.8e-5
	SLC30A7	zinc	8.8e-5
Mbuti	*SGK1*	selenium	6.77e-5
	SGCD	selenium	8.72e-5
	ANO3	chloride	0.00012945
	EEFSEC	selenium	0.00020261
	KCNMA1	calcium, potassium	0.00020261
Biaka	*DGKD*	calcium	1.97e-5
	NEDD4	sodium	3.93e-5
	SKG1	selenium	4.81e-5
	TMPRSS6	iron	7.18e-5
	SGCD	selenium	8.77e-5
Yoruba	*SELENOM*	selenium	5.87e-6
	SLC12A1	sodium, chloride, potassium	1.33e-5
	SCAMP5	zinc	4.17e-5
	CNNM2	magnesium	5.12e-5
	VWA8	calcium	0.00011987
Mandenka	*LRP8*	selenium	1.04e-5
	MTF1	zinc	1.04e-5
	TRPM2	calcium	2.11e-5
	SLC8A1	calcium	7.52e-5
	ANO3	chloride	7.52e-5
Mozabite	*CA1*	zinc	7.15E-06
	ATP2B2	calcium	1.38E-05
	FGFR2	magnesium	2.50E-05
	TPO	iodine	2.50E-05
	SLC12A1	sodium, chloride, potassium	2.54E-05
Palestinian	*THRB*	iodine	3.23E-06
	SLC39A11	zinc	1.48E-05
	CLCN3	chloride	2.62E-05
	PRKG1	selenium	6.83E-05
	SLC4A5	sodium	9.26E-05

	GALNT3	phosphorus	9.26E-05
Druze	*SLC12A1*	sodium, chloride, potassium	3.61E-05
	FGFR2	magnesium	6.67E-05
	GPHN	molybdenum	9.60E-05
	WDR45	iron	0.00010661
	SHROOM3	magnesium	0.00021632
Bedouin	*COMMD1*	copper	2.38E-05
	EPAS1	iron	0.00013935
	HBS1L	iron	0.0002138
	PRKG1	selenium	0.00022585
	TRPM6	magnesium	0.00022585
Adygei	*C19orf12*	iron	2.01E-05
	AKAP6	selenium	4.62E-05
	CFAP251	iron	9.28E-05
	SLC30A8	zinc	0.00015855
	ARHGEF3	iron	0.00015855
BergamoItalian-Tuscan	*SCNN1G*	sodium, potassium	1.11E-05
	SGK1	selenium	1.20E-05
	SLC8A1	calcium	4.48E-05
	SLC30A8	zinc	4.48E-05
	SLCO1C1	iodine	7.78E-05
Sardinian	*ATP2B2*	calcium	2.10E-07
	THRB	iron	2.94E-05
	SECISBP2	selenium, iodine	3.27E-05
	SLC8A1	calcium	0.00012589
	VWA8	selenium	0.0002822
	SOD1	copper	0.0002822
Basque	*HIF1A*	iron	2.43E-06
	ARHGEF3	iron	2.47E-05
	TXNRD3	selenium	4.79E-05
	EEFSEC	selenium	5.59E-05
	SLC34A2	phosphorus	8.55E-05
French	*SCNN1D*	sodium, potassium	1.87E-06
	ANO3	chloride	4.56E-05
	ATP2B2	calcium	7.51E-05
	SLC39A11	zinc	9.13E-05
	SLC12A1	sodium, chloride, potassium	0.00022405
Orcadian	*GPHN*	molybdenum	9.83E-05
	SLC5A5	sodium, iodine	9.83E-05
	SLC39A11	zinc	0.00011567
	CYP24A1	calcium	0.00011567
	FTMT	iron	0.00012535
Russian	*SLC4A5*	sodium	3.83E-06
	SCNN1D	sodium, potassium	6.81E-06
	SLC30A1	zinc	1.32E-05
	KCNMA1	calcium, potassium	0.000137
	SLC39A11	zinc	0.00015104
Makrani	*SLC39A11*	zinc	1.40E-06
	GPx2	selenium	9.61E-06
	SLC8A1	calcium	2.40E-05
	SLC39A12	zinc	2.40E-05
	ATP2B2	calcium	4.97E-05
Sindhi	*HIF1A*	iron	2.28E-05
	HSD11B2	iron	9.68E-05
	SLC39A11	zinc	0.00015917
	ANO6	chloride	0.00027978
	SLC30A9	zinc	0.00031178
Balochi	*PDE7B*	phosphorus	1.20E-05
	SLC39A11	zinc	1.55E-05
	SLC4A5	sodium	2.23E-05
	HSD11B2	iron	9.32E-05
	PRKG1	selenium	0.000145
Brahui	*MECOM*	magnesium	1.23E-06
	SLC4A5	sodium	1.08E-05
	SLC16A2	iodine	1.51E-05
	FTMT	iron	1.92E-05
	GPx2	selenium	0.00010148
	SGK1	selenium	0.00010148
Hazara	*PRKG1*	selenium	5.56E-05
	ARSB	selenium	0.00017433
	SELENOP	selenium	0.00017433
	TXNRD1	selenium	0.00017433
	LRP2	selenium	0.00034417

	CLIC5	chloride	0.00034417
Pathan	*ANO3*	chloride	1.76E-05
	ATP2B2	calcium	2.08E-05
	SELENOP	selenium	4.09E-05
	SLC12A1	sodium, chloride, potassium	5.81E-05
	SLC30A7	zinc	9.46E-05
Burusho	*AKR7L*	selenium	9.83E-06
	FGFR2	magnesium	4.75E-05
	ATP2B2	calcium	9.35E-05
	SLC39A11	zinc	0.00012304
	SLCO1C1	iodine	0.00012304
Kalash	*SLC39A10*	zinc, magnesium, manganese	1.21E-05
	STK39	sodium	1.97E-05
	SLC12A3	calcium, magnesium, potassium	2.21E-05
	PRKG1	selenium	0.00014809
	MECOM	magnesium	0.00014809
	BSND	chloride	0.00014809
Uygur	*FXYD2*	magnesium	2.80E-06
	SLC40A1	iron	1.62E-05
	ATP2B2	calcium	5.23E-05
	PLA2G6	iron	7.10E-05
	SLC30A9	zinc	0.0002744
	SLC8A1	calcium	0.0002744
Xibo-Mongolian	*SLC8A3*	calcium	6.74E-05
	KCNMA1	calcium, potassium	0.00016803
	PRKG1	selenium	0.0002743
	ANO3	chloride	0.0002743
	MT1H/F/G	zinc	0.0002743
	SELENOP	selenium	0.0002743
Oroqen-Hezhen-Daur	*SLC40A1*	iron	0.00023563
	PRKG1	selenium	0.00023563
	ARSB	selenium	0.00023563
	AKAP6	selenium	0.00023563
	ARHGEF3	iron	0.00023563
	SLCO1C1	iodine	0.00023563
Yakut	*FTMT*	iron	3.37E-06
	KCNMA1	calcium, potassium	1.46E-05
	AKAP6	selenium	7.62E-05
	GPx7	selenium	0.00013104
	IL6R	zinc	0.00013508
Japanese	*GPR39*	zinc	7.51E-05
	LHFPL2	selenium	0.00015814
	SLC40A1	iron	0.0001792
	DGKD	calcium	0.00020117
	ATP2B2	selenium	0.0002387
Han	*PRKG1*	selenium	6.54E-05
	SCNN1D	sodium, potassium	7.33E-05
	TRIP4	iodine	0.00013754
	SLCO1C1	iodine	0.00013754
	CFTR	chloride	0.00013754
NorthernHan-Tu	*SLC39A11*	zinc	3.90E-05
	SLC8A3	calcium	3.90E-05
	KCNMA1	calcium, potassium	3.90E-05
	PRKG1	selenium	0.00015139
	PDE7B	phosphorus	0.0001716
She-Miao-Tujia	*MLN*	phosphorus	4.27E-06
	ITPR3	phosphorus	1.93E-05
	SLC39A11	zinc	2.41E-05
	PRKG1	selenium	0.00013476
	FTMT	iron	0.00021508
Naxi-Yi	*PRKG1*	selenium	2.17E-05
	MOCS2	molybdenum	2.17E-05
	SLC39A11	zinc	8.43E-05
	SLC39A8	zinc, magnesium, manganese	0.00016511
	SLC8A3	calcium	0.00037186
	SELENOI	selenium	0.00037186
	IL6	zinc	0.00037186
	SLC30A10	zinc, magnesium, manganese	0.00037186
Dai-Lahu	*ITPR3*	phosphorus	4.43E-05
	SLC8A3	calcium	4.43E-05
	TRPM6	magnesium	6.09E-05
	ARHGEF3	iron	6.80E-05

	SLC8A1	calcium	7.39E-05
Pima	*ATP2B2*	iron	9.06E-06
	ITPR3	phosphorus	8.09E-05
	TRNAU1AP	selenium	0.00018445
	MLN	phosphorus	0.00053942
	LEMD2	phosphorus	0.00053942
Maya	*SLC8A1*	calcium	4.17E-05
	TRPM6	magnesium	5.25E-05
	ATP2B1	calcium	0.00015384
	ARL15	magnesium	0.00015794
	TSHR	iodine	0.00027492
	THRA	iodine	0.00027492
Surui-Karitiana	*AKAP6*	selenium	3.96E-05
	CLDN16	magnesium	0.00012466
	SLC39A10	zinc	0.00014355
	SLC39A8	zinc, magnesium, manganese	0.00020085
	MECOM	magnesium	0.00041299
	THRB	iodine	0.00041299
	SLC39A11	zinc	0.00041299
Papuan	*SLC8A1*	calcium	1.26E-05
	ATP2B2	calcium	1.97E-05
	DIO2	selenium, iodine	0.00027674
	HIF1A	iron	0.00027674
	HBS1L	iron	0.00027674
	SCNN1B	sodium, potassium	0.00027674
	LEMD2	phosphorus	0.00027674
Bougainville	*CLDN16*	magnesium	0.00021637
	SGCD	selenium	0.00021637
	NR3C2	sodium	0.0002471
	SLC39A11	zinc	0.000375
	ATP2B2	calcium	0.000375
	SLC30A6	zinc	0.000375
	CYP11B2	potassium	0.000375

Table S3.13: The five MAGs for each population with the strongest evidence for selection, as indicated by F_{ST} selection values*. When taking the only five MAGs would cut-off genes with the same significance value, more genes are given.*

Population	Gene	Micronutrient	Significance
San	*GALNT3*	phosphorus	3.50E-06
	SCNN1G	sodium, potassium	6.30E-06
	LRP8	selenium	2.15E-05
	LHFPL2	selenium	3.08E-05
	ANO7	chloride	7.75E-05
Bantu-speaking	*LHFPL2*	selenium	4.99E-06
	SLC12A1	sodium, chloride, potassium	5.06E-06
	KCNJ10	calcium, potassium	1.38E-05
	PRKG1	selenium	2.11E-05
	EEFSEC	selenium	2.53E-05
Mbuti	*TRIP4*	iodine	3.96E-05
	PDE7B	phosphorus	5.94E-05
	TRU-TCA2-1	selenium	6.63E-05
	LHFPL2	selenium	7.44E-05
	MECOM	magnesium	7.44E-05
Biaka	*SLC8A1*	calcium	3.05E-05
	ANO7	chloride	5.40E-05
	LHFPL2	selenium	8.33E-05
	ATP2B4	calcium	8.39E-05
	EEFSEC	selenium	0.00010422
Mandenka	*ATP2B2*	calcium	7.75E-08
	STK39	sodium	1.96E-05
	FTL	iron	1.99E-05
	HJV	iron	2.91E-05
	SLC8A1	calcium	3.00E-05
Mozabite	*SLC12A1*	sodium, chloride, potassium	7.35E-06
	EEFSEC	calcium	1.29E-05
	PDE7B	phosphorus	3.23E-05
	COMMD1	copper	7.20E-05
	ATP2B4	calcium	8.32E-05

Palestinian	*SLC12A1*	sodium, chloride, potassium	6.37E-07
	PDE7B	phosphorus	1.76E-05
	ARHGEF3	iron	2.55E-05
	SLC39A4	zinc	2.85E-05
	GPR39	zinc	3.72E-05
Druze	*SLC12A1*	sodium, chloride, potassium	2.97E-07
	PDE7B	phosphorus	8.16E-07
	SLC39A4	zinc	1.02E-05
	GPR39	zinc	3.23E-05
	MECOM	magnesium	5.36E-05
Bedouin	*SLC12A1*	sodium, chloride, potassium	1.25E-06
	PDE7B	phosphorus	1.37E-05
	SLC4A5	sodium	1.81E-05
	ARHGEF3	iron	3.43E-05
	GPR39	zinc	5.59E-05
Adygei	*SLC12A1*	sodium, chloride, potassium	2.01E-06
	PDE7B	phosphorus	2.05E-05
	SLC39A4	zinc	2.30E-05
	FTMT	iron	7.54E-05
	EEFSEC	selenium	0.00010962
BergamoItalian-Tuscan	*SLC12A1*	sodium, chloride, potassium	1.58E-06
	PDE7B	phosphorus	2.92E-06
	SLC39A4	zinc	8.84E-06
	SGCD	selenium	1.07E-05
	GPR39	zinc	8.94E-05
Sardinian	*PDE7B*	phosphorus	7.02E-07
	SLC12A1	sodium, chloride, potassium	6.86E-06
	SLC39A4	zinc	1.33E-05
	GPR39	zinc	5.57E-05
	MECOM	magnesium	6.35E-05
Basque	*SLC12A1*	sodium, chloride, potassium	2.00E-06
	PDE7B	phosphorus	2.00E-06
	SLC4A5	sodium	9.52E-06
	SLC39A4	zinc	1.41E-05
	GPR39	zinc	0.00013401
French	*SLC12A1*	sodium, chloride, potassium	7.65E-07
	PDE7B	phosphorus	4.28E-06
	SLC39A4	zinc	5.58E-06
	AQP6	chloride	3.41E-05
	GPR39	zinc	5.35E-05
Orcadian	*PDE7B*	phosphorus	2.24E-06
	SLC12A1	sodium, chloride, potassium	4.80E-06
	SLC39A4	zinc	3.19E-05
	ARSB	selenium	7.92E-05
	SCNN1A	sodium, potassium	9.97E-05
	SLC30A10	zinc, magnesium, manganese	9.97E-05
Russian	*SLC12A1*	sodium, chloride, potassium	1.40E-06
	PDE7B	phosphorus	5.27E-06
	SLC39A4	zinc	6.43E-06
	SLC4A5	sodium	1.89E-05
	SELENOS	selenium	4.01E-05
Makrani	*SLC39A4*	zinc	3.95E-06
	PDE7B	phosphorus	6.15E-06
	SLC12A1	sodium, chloride, potassium	7.37E-06
	SGCD	selenium	1.18E-05
	SGK1	selenium	3.59E-05
Sindhi	*PDE7B*	phosphorus	9.11E-06
	SLC39A4	zinc	1.03E-05
	CLCNKB	chloride	2.19E-05
	SLC39A11	zinc	2.99E-05
	GPR39	zinc	5.83E-05
Balochi	*SLC12A1*	sodium, chloride, potassium	1.76E-06
	SGK1	selenium	5.44E-06
	PDE7B	phosphorus	7.28E-06
	ARHGEF3	iron	2.34E-05
	SLC39A4	zinc	2.89E-05
Brahui	*SLC12A1*	sodium, chloride, potassium	1.07E-06
	PDE7B	phosphorus	1.53E-06
	SLC39A4	zinc	7.98E-06
	GPR39	zinc	2.63E-05
	SLC4A5	sodium	4.93E-05
Hazara	*SLC39A4*	zinc	7.24E-06
	SELENOS	selenium	1.86E-05

	SLC30A9	zinc	2.23E-05
	GPR39	zinc	9.24E-05
	MTF2	zinc	9.77E-05
Pathan	*SLC39A4*	zinc	5.47E-06
	SLC12A1	sodium, chloride, potassium	1.12E-05
	GPR39	zinc	3.47E-05
	PDE7B	phosphorus	4.92E-05
	SLC30A9	zinc	6.08E-05
Burusho	*SLC12A1*	sodium, chloride, potassium	6.07E-06
	SLC39A4	zinc	8.76E-06
	PDE7B	phosphorus	2.74E-05
	SLC30A9	zinc	5.38E-05
	GPR39	zinc	5.78E-05
Kalash	*SLC12A1*	sodium, chloride, potassium	2.56E-06
	SLC39A4	zinc	1.72E-05
	PDE7B	phosphorus	5.19E-05
	SLC39A11	zinc	9.98E-05
	GPR39	zinc	0.00013799
	EEFSEC	selenium	0.00013799
Uygur	*SLC39A4*	zinc	5.66E-05
	DCDC1	magnesium	0.00012776
	PDE7B	phosphorus	0.0001421
	SLCO1C1	iodine	0.00019682
	CA3	zinc	0.00019682
	PRKG1	selenium	0.00019889
Xibo-Mongolian	*PRKG1*	selenium	1.00E-05
	SLC30A9	zinc	2.66E-05
	SEPHS2	selenium	2.66E-05
	SLC39A4	zinc	5.49E-05
	HSD11B2	iron	6.97E-05
Oroqen-Hezhen-Daur	*SLC30A9*	zinc	1.51E-05
	SLC39A4	zinc	3.99E-05
	HSD11B2	iron	6.38E-05
	PDE7B	phosphorus	0.00012113
	KCNMA1	calcium, potassium	0.00015128
Yakut	*SLC30A9*	zinc	1.71E-05
	SLC39A4	zinc	3.57E-05
	PRKG1	selenium	0.00010185
	ANO5	chloride	0.00018071
	DIO1	selenium, iodine	0.0002481
Japanese	*SLC39A4*	zinc	6.69E-05
	RHOA	iron	6.91E-05
	SELENOW	selenium	6.91E-05
	SLC30A9	zinc	0.00014582
	CLCNKB	chloride	0.00017725
Han	*SLC30A9*	zinc	3.55E-06
	SLC39A4	zinc	4.34E-05
	RHOA	iron	8.62E-05
	CLCNKB	chloride	0.00013356
	ITPR3	phosphorus	0.00016139
NorthernHan-Tu	*SLC39A4*	zinc	5.98E-05
	SLC30A9	zinc	7.01E-05
	RHOA	iron	9.99E-05
	PRKG1	selenium	0.00011442
	SEPHS2	selenium	0.00020983
She-Miao-Tujia	*RHOA*	iron	1.38E-05
	SLC30A9	zinc	2.05E-05
	PRKG1	selenium	2.05E-05
	SLC39A4	zinc	4.75E-05
	CLCNKB	chloride	0.00014865
Naxi-Yi	*SLC39A4*	zinc	9.03E-05
	ITPR3	phosphorus	9.03E-05
	SLC30A9	zinc	0.00010709
	SEPHS2	selenium	0.00010969
	PRKG1	selenium	0.0001864
Dai-Lahu	*SLC30A9*	zinc	2.12E-05
	FTMT	iron	4.85E-05
	PDE7B	phosphorus	5.98E-05
	SLC39A4	zinc	7.82E-05
	SEPHS1	selenium	0.00013178
Pima	*PDE7B*	phosphorus	0.00010471
	GPx3	selenium	0.00010471
	CLCN3	chloride	0.00010471

	RHOA	iron	0.00010471
	SLC40A1	iron	0.00010471
	STAT3	zinc	0.00010471
	SLC39A11	zinc	0.00010471
	SLC30A2	zinc	0.00010471
	SELENON	selenium	0.00010471
Maya	*FGFR2*	magnesium	2.29E-05
	PDE7B	phosphorus	3.53E-05
	TMPRSS6	iron	4.90E-05
	SLC30A9	zinc	8.30E-05
	GPx3	selenium	0.00014139
Surui-Karitiana	*PDE7B*	phosphorus	0.00012002
	GPx3	selenium	0.00012002
	THRB	iodine	0.00012002
	SGCD	selenium	0.00012002
	SCNN1B	sodium, potassium	0.00012002
	SLC30A9	zinc	0.00012002
Papuan	*ACO1*	chloride	5.36E-05
	SGCD	selenium	5.36E-05
	TMPRSS6	iron	9.64E-05
	SLC39A11	zinc	9.64E-05
	SLC30A9	zinc	0.00025195
	CLCN3	chloride	0.00025195
Bougainville	*TFRC*	iron	3.80E-05
	DCDC1	magnesium	3.80E-05
	SLC30A9	zinc	8.23E-05
	TMPRSS6	iron	8.23E-05
	SCNN1G	sodium, potassium	8.23E-05
	SLC41A1	magnesium	8.23E-05

Table S3.14: MAG showing signatures in the 0.1% tail (for both Relate and F_{ST} selection values) for multiple populations.

Relate			***F_{ST}***		
Gene	**Micronutrient**	**Number of Repeats**	**Gene**	**Micronutrient**	**Number of Repeats**
Zinc	SLC39A11	27	Zinc	SLC39A4	37
Calcium	SLC8A1	26	Phosphorus	PDE7B	37
Calcium	ATP2B2	26	Zinc	GPR39	32
Selenium	PRKG1	24	Chloride	AQP6	32
Selenium	SGCD	21	Magnesium	DCDC1	29
Selenium	AKAP6	20	Selenium	EEFSEC	27
Selenium	KCNMA1	16	Zinc	SLC30A9	26
Chloride	ANO3	16	Calcium	ATP2B2	25
Phosphorus	PDE7B	15	Zinc	SLC39A11	24
Sodium, potassium	SCNN1D	13	Selenium	PRKG1	24
Sodium, potassium, chloride	SLC12A1	12	Iodine	TSHR	23
Magnesium	MECOM	12	Sodium, potassium, chloride	SLC12A1	22
Iron	FTMT	12	Selenium	SGCD	22
Iodine	THRB	12	Magnesium	MECOM	21
Iodine	SLCO1C1	11	Selenium	AKAP6	19
Sodium	SGK1	10	Iron	ARHGEF3	19
Iron	HIF1A	10	Selenium	SELENOS	18
Calcium	TRPM2	10	Sodium	SLC4A5	17
Calcium	SLC8A3	10	Chloride	ANO5	17
Selenium	SELENOF	9	Sodium	SGK1	16
Zinc	SLC30A8	8	Selenium, iodine	DIO1	16
Sodium	SLC4A5	8	Calcium	ATP2B4	16
Phosphorus	ITPR3	8	Molybdenum	GPHN	15
Zinc, manganese, magnesium	SLC30A10	8	Zinc, manganese	SLC39A14	14
Magnesium	FGFR2	8	Chloride	CLCNKB	14
Iron	ARHGEF3	8	Chloride	CFTR	13
Chloride	CFTR	8	Calcium	SLC8A2	13
Calcium	ATP2B4	8	Selenium	SEPHS2	12

Zinc	SLC30A9	7	Selenium	LRP8	12
Molybdenum	GPHN	7	Iodine	THRB	12
Iodine	SULT6B1	7	Calcium	SLC8A1	12
Chloride	ANO5	7	Zinc	SLC30A2	11
Calcium	VWA8	7	Sodium	HSD11B2	11
Selenium	LRP2	6	Selenium	SELENOI	11
Magnesium	TRPM6	6	Phosphorus	ITPR3	11
Iodine	SLC16A10	6	Magnesium	FGFR2	11
Chloride	ANO5	7	Iron	RHOA	11
Chloride	ANO6	6	Iron	FTMT	11
Zinc	GPR39	5	Zinc	MTF2	10
Sodium	NR3C2	5	Selenium, iodine	DIO2	10
Selenium	SELENOS	5	Chloride	CLCN3	10
Selenium	SELENOP	5	Selenium	ARSB	9
Selenium	SELENOI	5	Phosphorus	NBPF3	9
Selenium	LRP8	5	Iodine	SLCO1C1	9
Selenium	EEFSEC	5	Sodium, potassium	SCNN1D	8
Selenium	ARSB	5	Sodium, potassium	SCNN1B	8
Phosphorus	MLN	5	Selenium	SEPHS1	8
Magnesium	ARL15	5	Selenium	LHFPL2	8
Iron	SLC40A1	5	Selenium	KCNMA1	8
Chloride	CLCN6	5	Iron	TMPRSS6	8
Chloride	CLCN3	5	Iron	SLC40A1	8
Calcium	DGKD	5	Iron	EPAS1	8
Zinc	SLC39A3	4	Zinc	SLC39A8	7
Sodium	HSD11B2	4	Sodium, potassium	SCNN1A	7
Selenium	LHFPL2	4	Selenium	TXNRD2	7
Selenium	GPx2	4	Manganese, magnesium	SLC39A8	7
Potassium	HSD11B2	4	Iron	TMPRSS6	8
Phosphorus	GALNT3	4	Iron	HIF1A	7
Magnesium	SHROOM3	4	Selenium	GPx3	6
Magnesium	CLDN16	4	Copper	COMMD1	6
Iron	EPAS1	4	Chloride	ANO4	6
Copper	CCDC27	4	Calcium	CYP24A1	6
Chloride	CLCN7	4	Zinc, iron	SLC11A1	5
Zinc	SLC39A9	3	Zinc	CA1	5
Zinc	SLC39A12	3	Selenium	SELENOW	5
Zinc	SLC30A7	3	Selenium	SELENON	5
Zinc	SLC30A6	3	Iron	LTF	5
Zinc	MTF2	3	Iron	CFAP251	5
Zinc	CA3	3	Copper	CCDC27	5
Sodium	STK39	3	Chloride	ANO7	5
Sodium	SCNN1B	3	Zinc	STAT3	4
Selenium	TXNRD3	3	Zinc	SLC30A3	4
Selenium	TXNRD1	3	Zinc, manganese, magnesium	SLC30A10	4
Selenium, iodine	SECISBP2	3	Selenium	JMY	4
Selenium	SCLY	3	Selenium	DMGDH	4
Selenium, iodine	DIO2	3	Magnesium	TRPM6	4
Potassium	SCNN1B	3	Magnesium	SLC41A1	4
Phosphorus	SLC34A2	3	Magnesium	EGF	4
Phosphorus	LEMD2	3	Iron	LCN2	4
Iron	SLC48A1	3	Iron	ACO1	4
Iron	RHOA	3	Zinc	SCAMP5	3
Iron	PLA2G6	3	Zinc	IL6R	3
Iron	PANK2	3	Sodium	NR3C2	3
Iodine	TTR	3	Manganese	ATP2C1	3
Iodine	TRIP4	3	Magnesium	ARL15	3
Copper	COMMD1	3	Iron	TFRC	3
Calcium	GATA3	3	Iron	FTH1	3
Calcium	ATP2B1	3	Chloride	SLC12A2	3
Zinc	SLC39A8	2	Chloride	CLIC4	3
Zinc	SLC39A10	2	Chloride	BEST1	3
Zinc	MT1H	2	Chloride	ANO3	3
Zinc	MT1G	2	Zinc	SLC39A9	2
Zinc	MT1F	2	Zinc	SLC39A2	2
Zinc	MT1A	2	Zinc	SLC39A13	2
Zinc	IL6R	2	Zinc	SLC39A12	2
Sodium	SCN3B	2	Zinc	SLC30A8	2
Sodium	NEDD4	2	Zinc	CAR13	2
Selenium	SEPSECS	2	Zinc	CA3	2
Selenium	GPx7	2	Sodium	STK39	2
Selenium	ELAVL1	2	Selenium	TRU-TCA2-1	2
Potassium	CYP11B2	2	Selenium	GPx2	2

Phosphorus	FGF6	2
Molybdenum	MOCS2	2
Manganese, magnesium	SLC39A8	2
Manganese	ATP2C1	2
Magnesium	CNNM2	2
Iron	TMPRSS6	2
Iron	TFRC	2
Iron	SLC17A1	2
Iron	MYB	2
Iron	LCN2	2
Iron	HBS1L	2
Iron	FA2H	2
Iron	CFAP251	2
Iron	C19orf12	2
Iron	ACO1	2
Iodine	TSHR	2
Iodine	THRA	2
Iodine	IYD	2
Copper	ATP7B	2
Copper	ATP7A	2
Chloride	SLC12A2	2
Chloride	CLIC5	2
Chloride	CLCNKB	2
Chloride	BSND	2
Chloride	ANO4	2
Calcium	SLC8A2	2
Calcium	CYP24A1	2
Chloride	ANO4	2
Calcium	SLC8A2	2
Calcium	CYP24A1	2
Calcium	SLC8A2	2
Calcium	CYP24A1	2
Potassium	SCNN1G	2
Phosphorus	GALNT3	2
Phosphorus	CASR	2
Magnesium	SHROOM3	2
Magnesium	KCNA1	2
Magnesium	CASR	2
Iron	STEAP3	2
Iron	HJV	2
Iron	FTL	2
Iodine	TRIP4	2
Iodine	THRA	2
Iodine	SULT6B1	2
Chloride	CLCN6	2
Calcium	TRPM2	2
Calcium	SLC8A3	2
Calcium	KCNJ10	2
Calcium	GCKR	2
Calcium	CASR	2

Figures

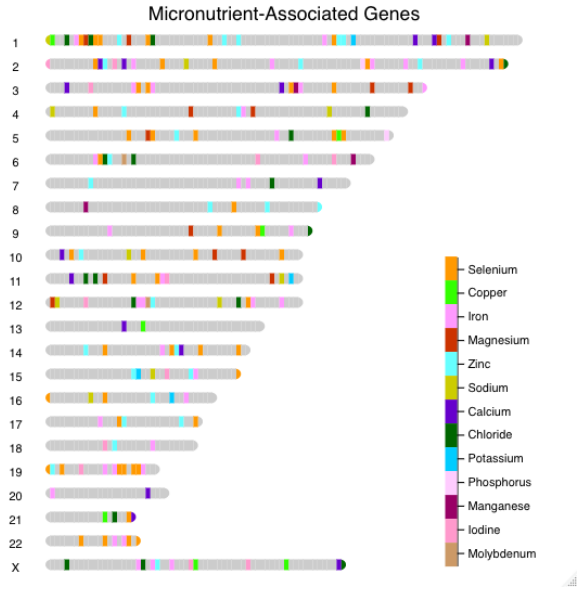


***Figure S3.1**: **Distribution of all micronutrient-associated genes along the human genome**. Broadly randomly distributed with any overlaps given in **Table S4.3.2**.*

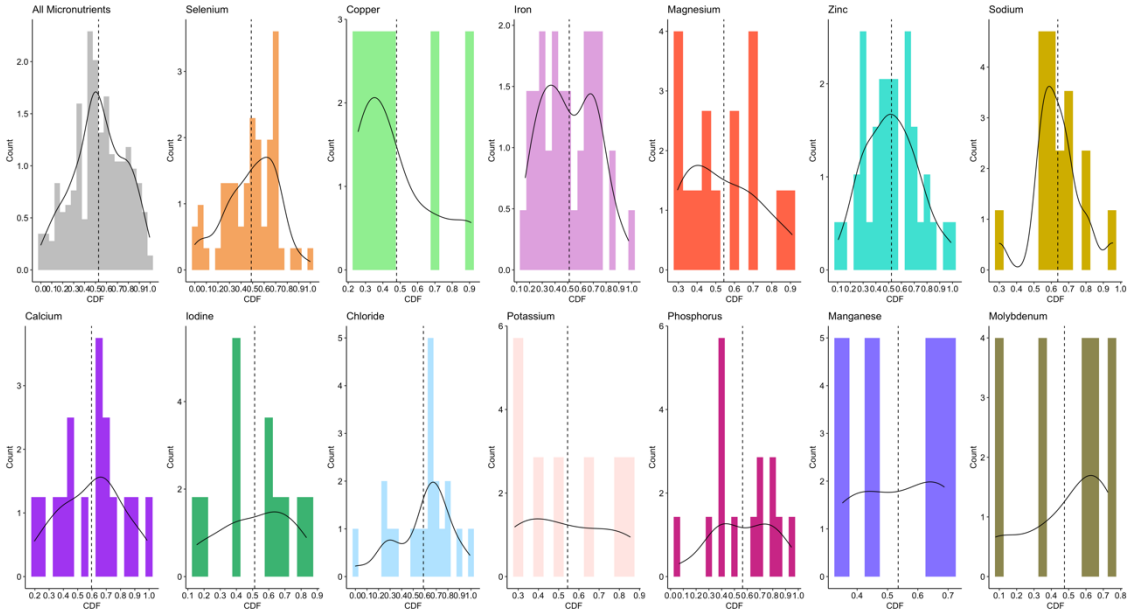


***Figure S3.2*: *Distribution of the calculated CDF value*.** *As drawn from the distribution formed from generated neutral gene regions and includes the mean of cumulative density function (CDF) position.*

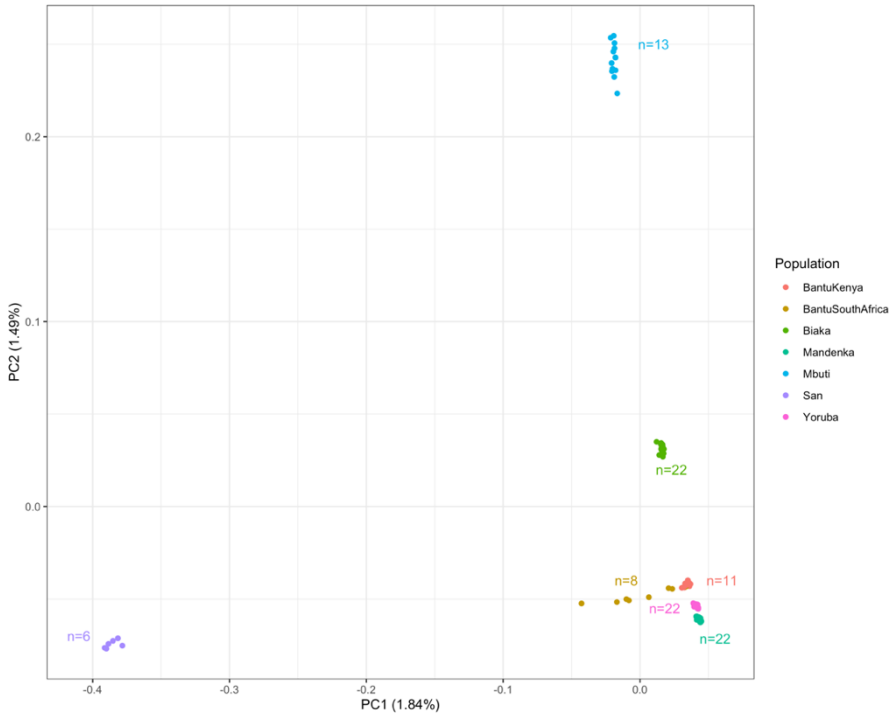


***Figure S3.3: Principal component analysis of African individuals*.** *From the (Bergström et al., 2020), showing PC1 and PC2.*

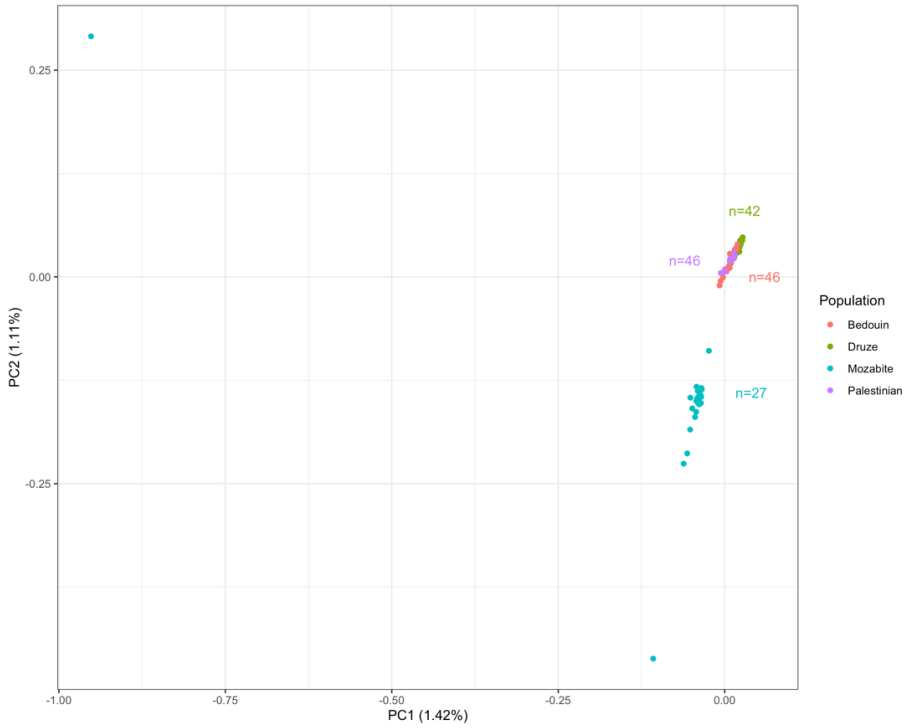


Figure S3.4: Principal component analysis of Middle-eastern individuals*. From (Bergström et al., 2020), showing PC1 and PC2, having removed outlier individuals*

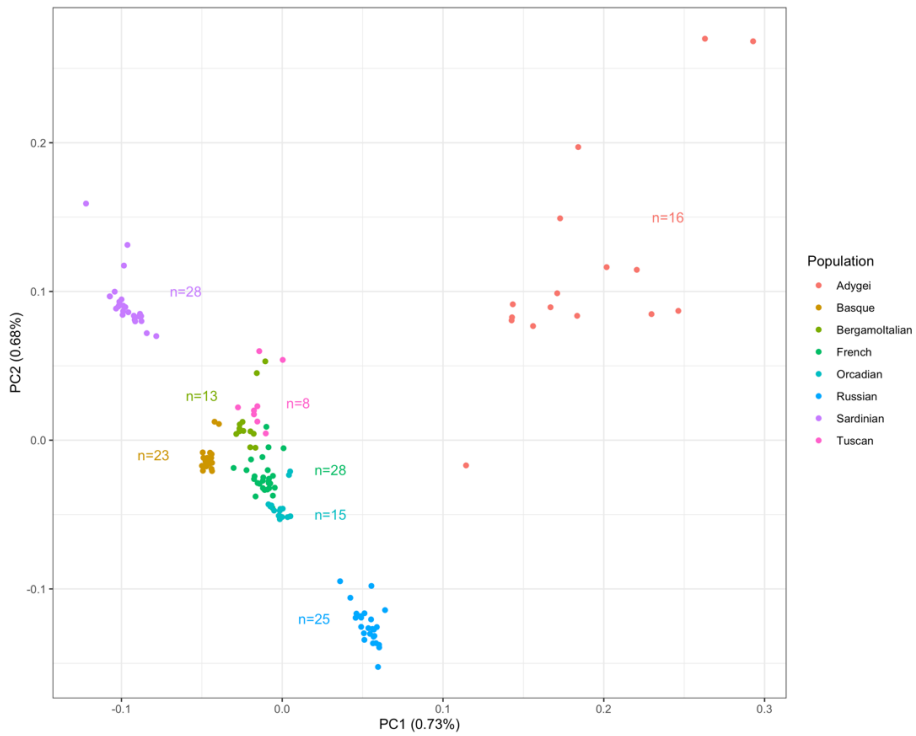


Figure S3.5: Principal component analysis of European individuals. *From the (Bergström et al., 2020), showing PC1 and PC2, having removed outlier individuals*.

Figure S3.6: Principal component analysis of Central-South Asian individuals. *From the (Bergström et al., 2020), showing PC1 and PC2, having removed outlier individuals*

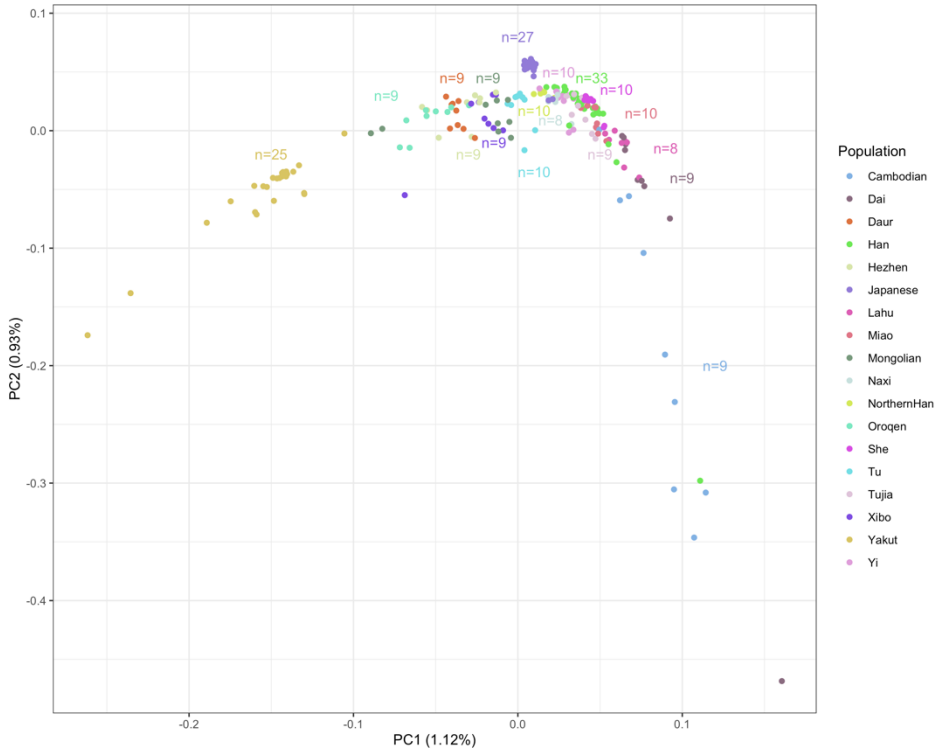


Figure S3.7: Principal component analysis of East Asian individuals*. From the (Bergström et al., 2020), showing PC1 and PC2*

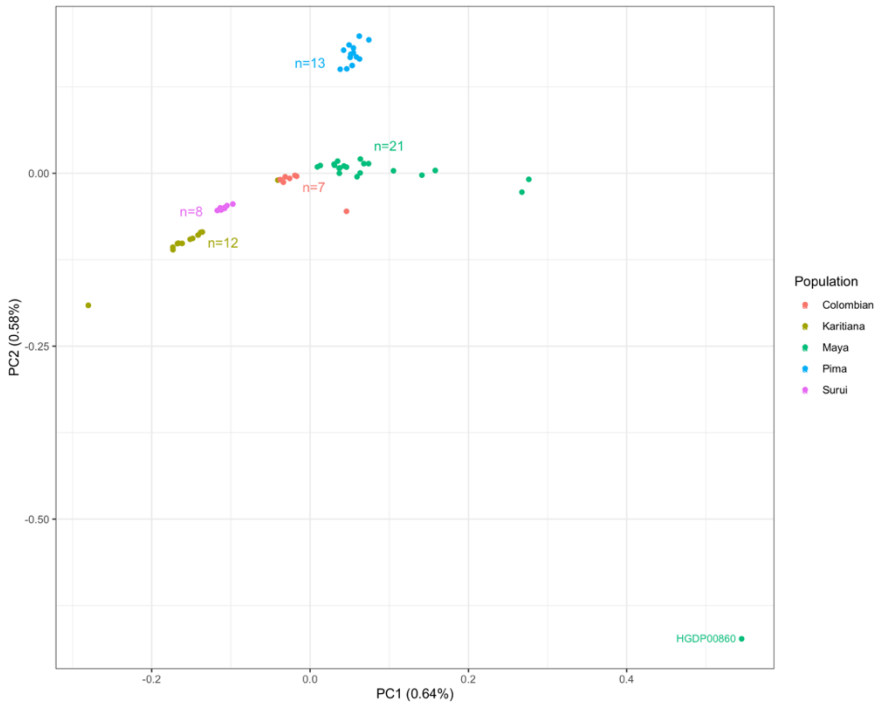


***Figure S3.8: Principal component analysis of American individuals**. From the (Bergström et al., 2020), showing PC1 and PC2, having removed outlier individuals*

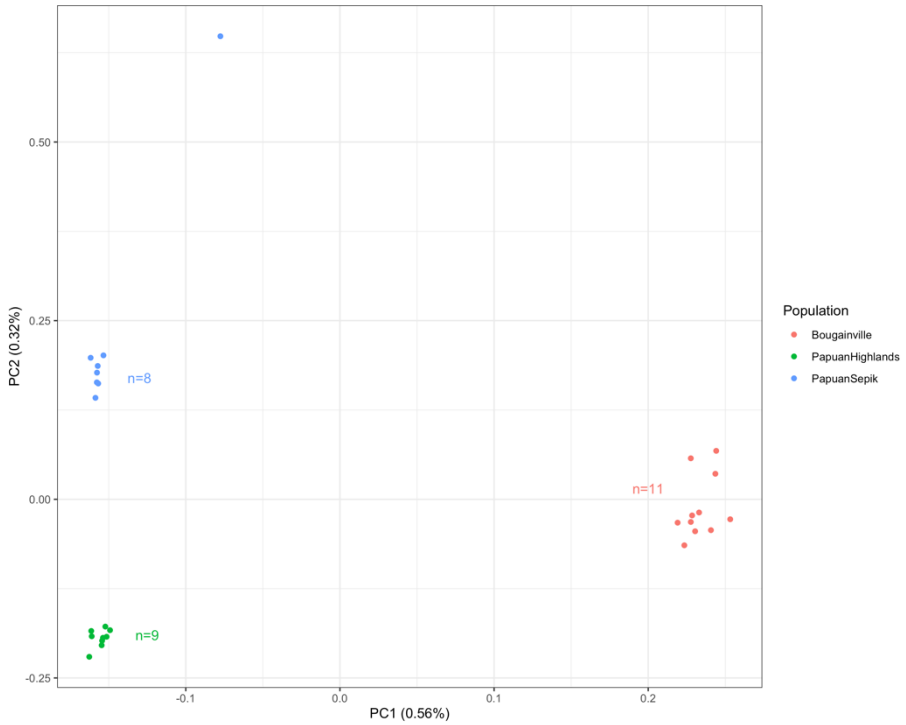


***Figure S3.9: Principal component analysis of Oceanic individuals**. From the (Bergström et al., 2020), showing PC1 and PC2.*

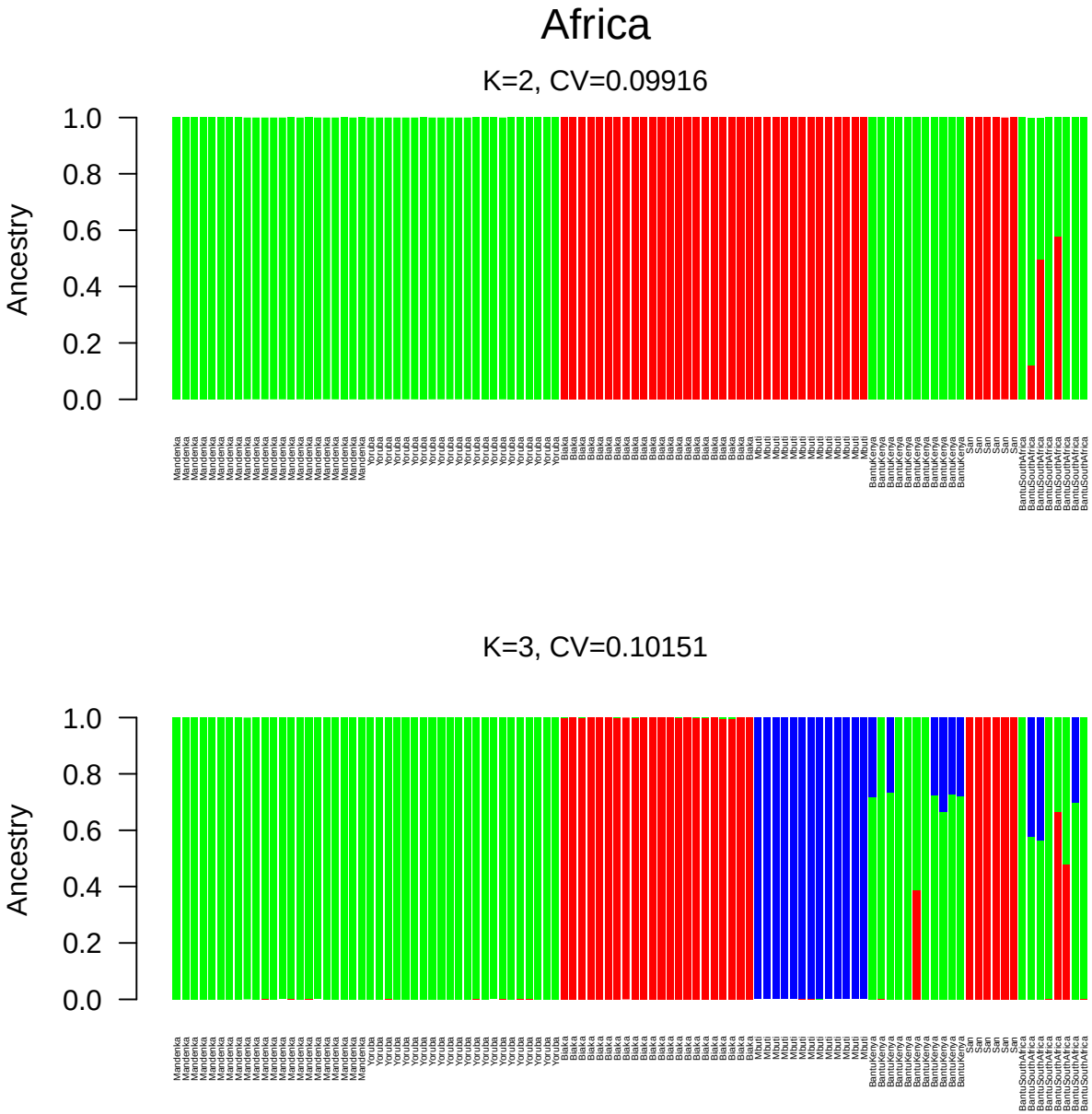


Figure S3.10: Admixture analysis of African individuals*. From the (Bergström et al., 2020) for 2 and 3 k clusters.*

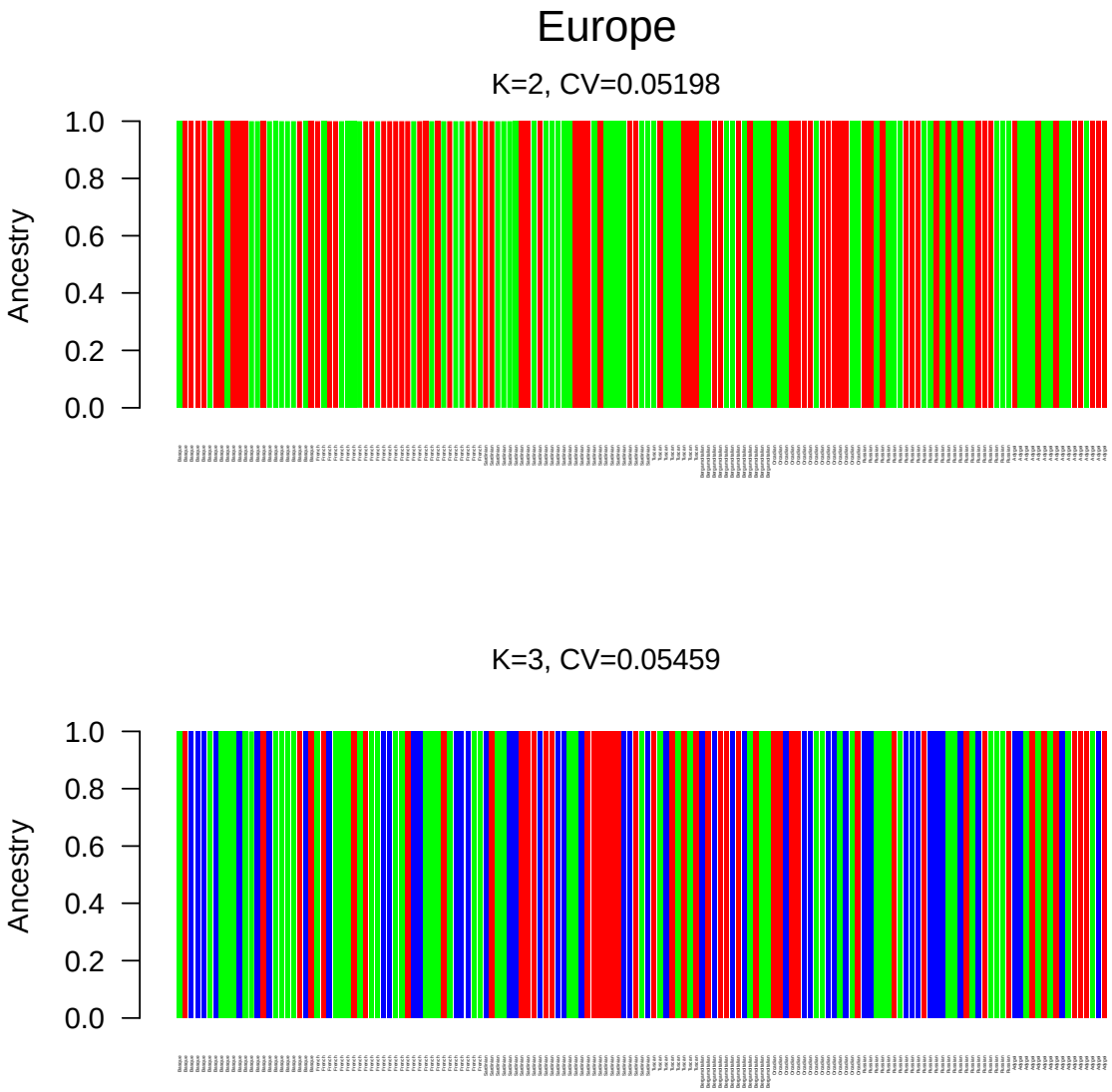


Figure S3.11: Admixture analysis of European individuals*. From the (Bergström et al., 2020) for 2 and 3 k clusters.*

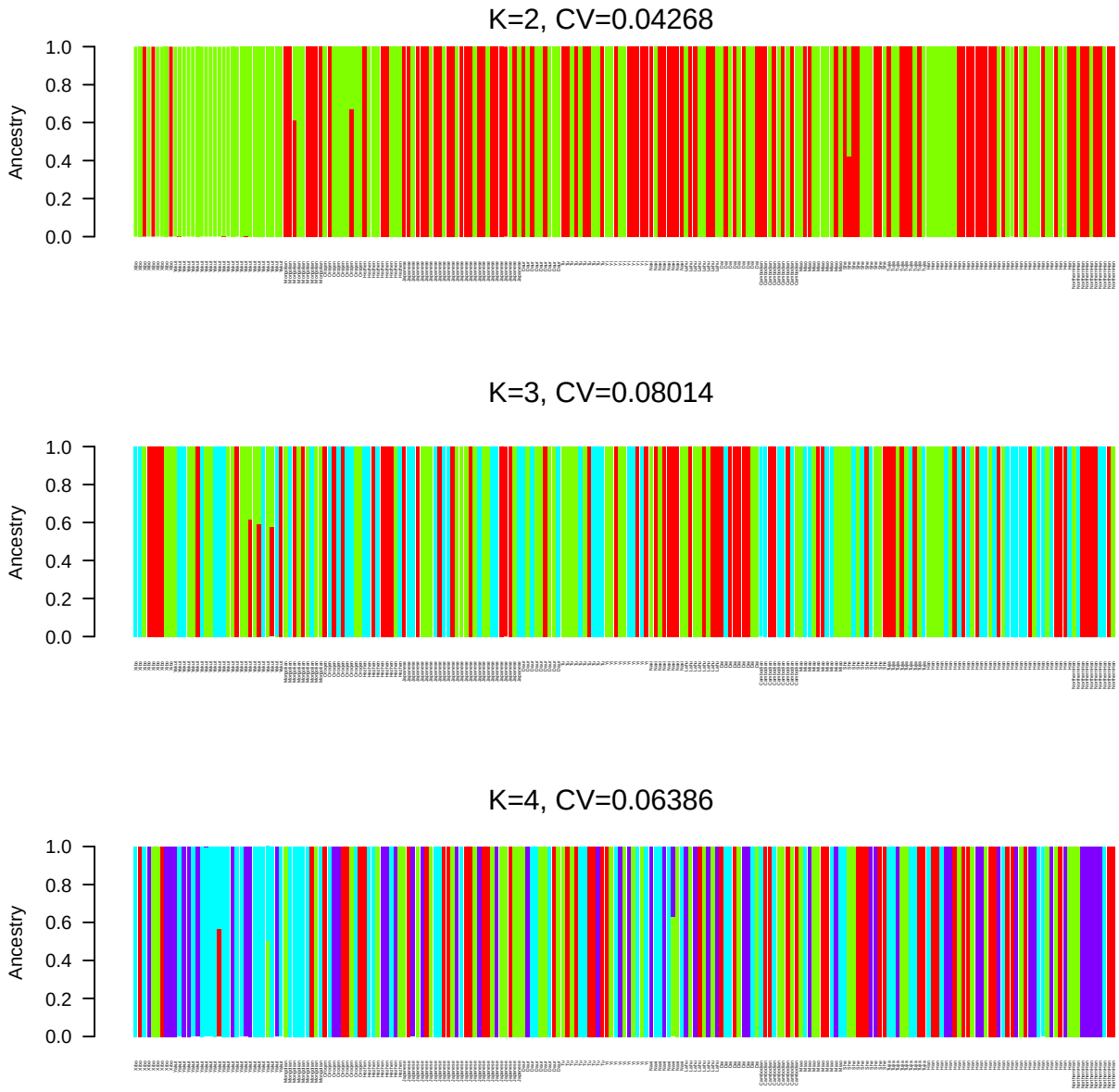


Figure S4.3.12: Admixture analysis of East Asian individuals*. From the (Bergstrom et al, 2020) for 2, 3 and 4 k clusters.*

Chapter 4: Supplementary Material

Tables

***Table S4.1**: All micronutrient-associated genes used in this study associated with the uptake, metabolism or regulation of selenium, zinc, iron, iodine and calcium. When genes are associated with multiple micronutrients, this is given in the "Other Associations" column. Genes removed following the positive mask (Bergström et al., 2020) indicated in the "Removed During Pruning" column. Gene regions as taken from ensemble (Yates et al., 2020) and suggested from the literature ("Reference").*

Micronutrient	Gene	Chromosome	Gene Start	Gene End	Other Associations	Removed During Pruning	Ref
Selenium	GPx1	3	49357176	49358358			(White et al., 2015)
Selenium	GPx2	14	64939152	64942905			(White et al., 2015)
Selenium	GPx3	5	151020438	151028992			(White et al., 2015)
Selenium	GPx4	19	1103926	1106791			(White et al., 2015)
Selenium	GPx6	6	28503296	28528215			(White et al., 2015)
Selenium	DIO1	1	53891239	53911086	iodine		(White et al., 2015)
Selenium	DIO2	14	80197526	80387757	iodine		(White et al., 2015)
Selenium	DIO3	14	101561351	101563452	iodine		(White et al., 2015)
Selenium	SELENOF	1	86862445	86914424			(White et al., 2015)
Selenium	SELENOH	11	57741250	57743554			(White et al., 2015)
Selenium	SELENOI	2	26308547	26395891			(White et al., 2015)
Selenium	SELENOK	3	53884417	53891962			(White et al., 2015)
Selenium	SELENOM	22	31104772	31120069			(White et al., 2015)
Selenium	SELENON	1	25800176	25818221			(White et al., 2015)
Selenium	SELENOO	22	50200979	50217616			(White et al., 2015)
Selenium	SELENOT	3	150602875	150630445			(White et al., 2015)
Selenium	SELENOV	19	39515113	39520686			(White et al., 2015)
Selenium	SELENOW	19	47778585	47784686			(White et al., 2015)
Selenium	MSRB1	16	1938210	1943326			(White et al., 2015)
Selenium	TXNRD1	12	104215779	104350307			(White et al., 2015)
Selenium	TXNRD2	22	19875517	19941820			(White et al., 2015)
Selenium	TXNRD3	3	126607059	126655124			(White et al., 2015)
Selenium	GPx5	6	28525881	28534955		Yes	(White et al., 2015)
Selenium	GPx7	1	52602371	52609051			(White et al., 2015)

Selenium	GPx8	5	55160167	55167297			(White et al., 2015)
Selenium	SELENOP	5	42799880	42887392			(White et al., 2015)
Selenium	LRP8	1	53242364	53328469			(White et al., 2015)
Selenium	LRP2	2	169127109	169362534			(White et al., 2015)
Selenium	SCLY	2	238060924	238099413			(White et) al., 2015)
Selenium	SELENBP1	1	151364304	151372707			(White et al., 2015)
Selenium	PSTK	10	122954381	122997513			(White et al., 2015)
Selenium	SEPSECS	4	25120014	25160449			(White et al., 2015)
Selenium	SARS2	19	38915266	38930763			(White et al., 2015)
Selenium	TRU-TCA1-1	19	45478602	45478687			(White et al., 2015)
Selenium	TRU-TCA2-1	22	44150657	44150742			(White et al., 2015)
Selenium	TRU-TCA3-1	17	40117300	40117373			(White et al., 2015)
Selenium	CELF1	11	47465933	47565569			(White et al., 2015)
Selenium	EEFSEC	3	128153481	128408646			(White et al., 2015)
Selenium	EIF4A3	17	80134369	80147151			(White et al., 2015)
Selenium	ELAVL1	19	7958573	8005659			(White et al., 2015)
Selenium	RPL30	8	98024851	98046469			(White et al., 2015)
Selenium	SECISBP2	9	89318500	89359663	iodine		(White et al., 2015)
Selenium	SEPHS1	10	13317428	13348298			(White et al., 2015)
Selenium	TRNAU1AP	1	28553085	28578545			(White et al., 2015)
Selenium	XPO1	2	61477849	61538626			(White et al., 2015)
Selenium	AKAP6	14	32329298	32837684			(Engelken et al., 2016)
Selenium	FABP1	2	88122982	88128062			(Engelken et al., 2016)
Selenium	KCNMA1	10	76869601	77638369			(Engelken et al., 2016)
Selenium	PRKG1	10	50990888	52298423			(Engelken et al., 2016)
Selenium	SELENOS	15	101270817	101277500			(Engelken et al., 2016)
Selenium	SEPHS2	16	30443631	30445874			(Engelken et al., 2016)
Selenium	SGCD	5	155870344	156767788			(Engelken et al., 2016)
Selenium	TXN	9	110243810	110256507			(Engelken et al., 2016)
Selenium	AKR7L	1	19265982	19274194			(Wishart et al., 2007)
Selenium	CBS	21	43053191	43076943		Yes	(Dib et al., 2019)
Selenium	ARSB	5	78777209	78986087			(Dib et al., 2019)
Selenium	LHFPL2	5	78485215	78770021			(Dib et al., 2019)
Selenium	DMGDH	5	78997564	79236038			(Dib et al., 2019)
Selenium	BHMT2	5	79069767	79090069			(Dib et al., 2019)
Selenium	BHMT	5	79111809	79132288			(Dib et al., 2019)
Selenium	JMY	5	79236131	79327211			(Dib et al., 2019)

Iron	BDH2	4	103077592	103099870			(Engelken et al., 2016)
Iron	CYBRD1	2	171522247	171558129			(Engelken et al., 2016)
Iron	EPAS1	2	46293667	46386697			(Engelken et al., 2016)
Iron	FECH	18	57544377	57586702			(Engelken et al., 2016)
Iron	FTH1	11	61959718	61967634			(Engelken et al., 2016)
Iron	FTL	19	48965309	48966879			(Engelken et al., 2016)
Iron	HAMP	19	35280716	35285143			(Engelken et al., 2016)
Iron	HEPH	X	66162549	66268867			(Engelken et al., 2016)
Iron	HFE	6	26087281	26098343			(Engelken et al., 2016)
Iron	HJV	1	146017468	146036746			(Engelken et al., 2016)
Iron	HIF1A	14	61695513	61748259			(Engelken et al., 2016)
Iron	LTF	3	46435645	46485234			(Engelken et al., 2016)
Iron	RHOA	3	49359145	49412998			(Engelken et al., 2016)
Iron	SLC17A1	6	25782915	25832052			(Engelken et al., 2016)
Iron	SLC40A1	2	189560590	189583758			(Engelken et al., 2016)
Iron	STEAP3	2	119223831	119265652			(Engelken et al., 2016)
Iron	TF	3	133746040	133796641			(Engelken et al., 2016)
Iron	TFR2	7	100620416	100642779			(Engelken et al., 2016)
Iron	TFRC	3	196027183	196082096			(Engelken et al., 2016)
Iron	TMPRSS6	22	37065436	37109713			(Engelken et al., 2016)
Iron	ISCU	12	108562582	108569384			(Engelken et al., 2016)
Iron	LCN2	9	128149071	128153453			(Engelken et al., 2016)
Iron	FTMT	5	121851882	121852833			(Wishart et al., 2007)
Iron	ACO1	9	32384603	32454769			(Muckenthaler et al., 2008)
Iron	IREP2	15	78437431	78501453			(Muckenthaler et al., 2008)
Iron	ACO2	22	41469117	41528989			(Muckenthaler et al., 2008)
Iron	ALAS2	X	55009055	55030977			(Muckenthaler et al., 2008)
Iron	SLC46A1	17	28394642	28407197			(Muckenthaler et al., 2008)
Iron	SLC11A1	2	218382029	218396894	zinc		(Fishilevich et al., 2017)
Iron	SLC48A1	12	47753916	47782751			(Fishilevich et al., 2017)
Iron	SLC11A2	12	50979401	51028566			(Muckenthaler et al., 2008)
Iron	HBS1L	6	134960378	135103056			(Dib et al., 2019)
Iron	MYB	6	135181308	135219173			(Dib et al., 2019)
Iron	PIK3CG	7	106865278	106908980			(Dib et al., 2019)
Iron	CFAP251	12	121918592	122003927			(Dib et al., 2019)
Iron	ARHGEF3	3	56727418	57079329			(Dib et al., 2019)
Iron	TAOK1	17	29390464	29551904			(Dib et al., 2019)

Iron	CP	3	149162410	149221829			(Fishilevich et al., 2017)
Iron	PANK2	20	3888839	3929882			(Fishilevich et al., 2017)
Iron	PLA2G6	22	38111495	38214778			(Fishilevich et al., 2017)
Iron	C19orf12	19	29698886	29715789			(Fishilevich et al., 2017)
Iron	FA2H	16	74712955	74774831			(Fishilevich et al., 2017)
Iron	WDR45	X	49074433	49101170			(Fishilevich et al., 2017)
Iron	ATP13A2	1	16985958	17011928			(Fishilevich et al., 2017)
Zinc	GPR39	2	132416805	132646582			(Engelken et al., 2016)
Zinc	IL6	7	22725884	22732002			(Engelken et al., 2016)
Zinc	IL6R	1	154405193	154469450			(Engelken et al., 2016)
Zinc	MT1A	16	56638666	56640087			(Engelken et al., 2016)
Zinc	MT1E	16	56625475	56627112			(Engelken et al., 2016)
Zinc	MT1F	16	56657731	56660698			(Engelken et al., 2016)
Zinc	MT1G	16	56666730	56668065			(Engelken et al., 2016)
Zinc	MT1H	16	56669814	56671129			(Engelken et al., 2016)
Zinc	MT2A	16	56608584	56609497			(Engelken et al., 2016)
Zinc	MT4	16	56565073	56568957			(Engelken et al., 2016)
Zinc	MTF1	1	37809574	37859592			(Engelken et al., 2016)
Zinc	MTF2	1	93079235	93139079			(Engelken et al., 2016)
Zinc	SLC11A1	2	218382029	218396894	iron		(Fishilevich et al., 2017)
Zinc	SLC30A1	1	211571568	211579161			(Engelken et al., 2016)
Zinc	SLC30A2	1	26037252	26046118			(Engelken et al., 2016)
Zinc	SLC30A3	2	27253684	27275817			(Engelken et al., 2016)
Zinc	SLC30A4	15	45479606	45522755			(Engelken et al., 2016)
Zinc	SLC30A5	5	69093949	69131069			(Engelken et al., 2016)
Zinc	SLC30A6	2	32165841	32224379			(Engelken et al., 2016)
Zinc	SLC30A7	1	100896076	100981757			(Engelken et al., 2016)
Zinc	SLC30A8	8	116950273	117176714			(Engelken et al., 2016)
Zinc	SLC30A9	4	41990502	42090461			(Engelken et al., 2016)
Zinc	SLC39A1	1	153959099	153968184			(Engelken et al., 2016)
Zinc	SLC39A10	2	195575977	195737702			(Engelken et al., 2016)
Zinc	SLC39A11	17	72645949	73092712			(Engelken et al., 2016)
Zinc	SLC39A12	10	17951839	18043292			(Engelken et al., 2016)
Zinc	SLC39A13	11	47407132	47416496			(Engelken et al., 2016)
Zinc	SLC39A2	14	20999255	21001871			(Engelken et al., 2016)
Zinc	SLC39A3	19	2732204	2740028			(Engelken et al., 2016)
Zinc	SLC39A4	8	144409742	144416844			(Engelken et al., 2016)

Zinc	SLC39A5	12	56230049	56237846			(Engelken et al., 2016)
Zinc	SLC39A6	18	36108531	36129385			(Engelken et al., 2016)
Zinc	SLC39A7	6	33200445	33204439		Yes	(Engelken et al., 2016)
Zinc	SLC39A8	4	102251041	102431258			(Engelken et al., 2016)
Zinc	SLC39A9	14	69398015	69462390			(Engelken et al., 2016)
Zinc	STAT3	17	42313324	42388568			(Engelken et al., 2016)
Zinc	SLC30A10	1	219685427	219958647			(Dib et al., 2019)
Zinc	SLC39A14	8	22367249	22434129			(Horning et al., 2015)
Zinc	CA1	8	85327608	85379014			(Dib et al., 2019)
Zinc	CA2	8	85463968	85481493			(Dib et al., 2019)
Zinc	CA3	8	85373436	85449040			(Dib et al., 2019)
Zinc	CAR13	8	85220587	85284073			(Dib et al., 2019)
Zinc	SCAMP5	15	74957219	75021495			(Dib et al., 2019)
Zinc	KLF8	X	56232356	56291531			(Dib et al., 2019)
Zinc	ZXDA	X	57906708	57910820			(Dib et al., 2019)
Zinc	ZXDB	X	57591652	57597545			(Dib et al., 2019)
Calcium	TRPM2	21	44350163	44443081			(Engelken et al., 2016)
Calcium	TRPV5	7	142908101	142933746		Yes	(Kovacs et al., 2013)
Calcium	TRPV6	7	142871208	142885745		Yes	(Hughes et al., 2008)
Calcium	CASR	3	122183668	122291629			(Houillier, 2014)
Calcium	BSPRY	9	113349541	113371233			(Khanal & Nemere, 2008)
Calcium	RGS2	1	192809039	192812275			(Khanal & Nemere, 2008)
Calcium	SLC8A1	2	40097270	40611053			(Khanal & Nemere, 2008)
Calcium	SLC8A2	19	47428017	47471893			(Khanal & Nemere, 2008)
Calcium	SLC8A3	14	70044215	70189070			(Khanal & Nemere, 2008)
Calcium	ATP2B2	3	10324023	10708007			(Khanal & Nemere, 2008)
Calcium	ATP2B3	X	153517676	153582939			(Khanal & Nemere, 2008)
Calcium	ATP2B4	1	203626561	203744081			(Khanal & Nemere, 2008)
Calcium	PTH	11	13492054	13496181			(Khanal & Nemere, 2008)
Calcium	CYP24A1	20	54153446	54173986			(Dib et al., 2019)
Calcium	GATA3	10	8045378	8075198			(Dib et al., 2019)
Calcium	DGKD	2	233354507	233472104			(Dib et al., 2019)

Calcium	VWA8	13	41566835	41961120			(Dib et al., 2019)
Calcium	GCKR	2	27496839	27523684			(Dib et al., 2019)
Calcium	KCNJ10	1	159998651	160070160			(Jain et al., 2013)
Calcium	SLC12A3	16	56865207	56915850			(Jain et al., 2013)
Calcium	SLC34A1	5	177379235	177398848			(Chang & Anderson, 2017)
Calcium	SLC34A3	9	137230757	137236555			(Chang & Anderson, 2017)
Iodine	DIO1	1	53891239	53911086	selenium		(White et al., 2015)
Iodine	DIO2	14	80197526	80387757	selenium		(White et al., 2015)
Iodine	DIO3	14	101561351	101563452	selenium		(White et al., 2015)
Iodine	TRIP4	15	64387748	64455303			(Herráez et al., 2009)
Iodine	IYD	6	150368892	150405969			
Iodine	SLC5A5	19	17871945	17895174			(Engelken et al. 2016)
Iodine	SLC16A10	6	111087503	111231194			(The UniProt Consortium 2023)
Iodine	THRA	17	40058290	40093867			(The UniProt Consortium 2023)
Iodine	THRB	3	24117153	24495756			(The UniProt Consortium 2023)
Iodine	SLC16A2	X	74421493	74533917			(The UniProt Consortium 2023)
Iodine	TSHR	14	80954989	81146302			(The UniProt Consortium 2023)
Iodine	SLCO1C1	12	20695355	20753386			(The UniProt Consortium 2023)
Iodine	SECISBP2	9	89318500	89359663	selenium		(White et al. 2015)
Iodine	TPO	2	1374066	1543711			(Wishart et al. 2007)
Iodine	TTR	18	31557010	31599021			(Wishart et al. 2007)
Iodine	SERPINA7	X	106032435	106038727			(Wishart et al. 2007)
Iodine	SLC3A2	11	62856102	62888875			(Wishart et al. 2007)
Iodine	SULT6B1	2	37167820	37196598			(Wishart et al. 2007)

***Table S4.2**: **ZCSII-associated genes and their associated focal SNPs showing high repetition of selection signatures**. "Gene Repetition"; given for both Relate and F_{ST} selection values.*

Micronutrient	**Gene**	**Gene Repetition** ***Relate***	F_{ST}	**Focal SNP Position**
Zinc	*SLC39A4*	1	14	chr8:144414297
	GPR39	5	32	chr2:132638916
	SLC30A9	7	26	chr4:42004040 chr4:42031397

				chr4:42066213 chr4:42093983
	SLC39A11	27	24	chr17:73010373 chr17:72716374
	SLC39A14	1	14	chr8:22404076 chr8:22416174
Calcium	*ATP2B2*	26	25	chr3:10453703 chr3:10636328
	ATP2B4	8	16	chr1:203648263 chr1:203667951
	SLC8A2	2	13	chr19:47428756 chr19:47437107
Selenium	*EEFSEC*	5	27	chr3:128412869
	PRKG1	24	24	chr10:51576270 chr10:51471686
	SGCD	21	22	chr5:156708844 chr5:156057959
	AKAP6	20	19	chr14:32542441 chr14: 32446036 chr14: 32453376
	DIO1	0	16	chr1:53920598
Iron	*ARHGEF3*	8	19	chr3:56761998
Iodine	*TSHR*	2	23	chr14:80962759 chr14:81006112 chr14:81071140
	THRB	12	12	chr3: 24110895 chr3: 24342863

***Table S4.3*: Iron and Calcium-associated genes and their associated focal SNPs showing high repetition of selection signatures**. *"Gene Repetition"; given for both Relate and F_{ST} selection values.*

Micronutrient	**Gene**	**Gene Repetition**		**Focal SNP Position**
		Relate	F_{ST}	
Calcium	*ATP2B2*	26	25	chr3:10456514 chr3:10604833
	ATP2B4	8	16	chr1:203648263 chr1:203667951
	SLC8A1	26	12	chr2:40394610 chr2:40584510
	SLC8A2	2	13	chr19:47428756 chr19:47437107
	SLC8A3	10	2	chr14:70182346 chr14:70175561
Iron	*ARHGEF3*	8	19	chr3:56761998 chr3:57043874
	HIF1A	10	7	chr14:61687412 chr14:61709502 chr14:61741756
	FTMT	12	11	chr5:121846819 chr5:121853801
	SLC40A1	5	8	chr2:189577426 chr2:189591670

***Table S4.4**: The Zinc-associated genes within the 0.1% tail, as indicated by the **Relate selection values for each population**. Ordered by most significant.*

Region	Population	Gene	$Relate\ P-value$
Africa	Bantu-speaking	*SLC39A11*	3.33E-05
		SLC30A7	8.80E-05
		IL6R	0.00019916
		SLC30A8	0.00049756
		MT1A	0.00066041
	Biaka	*CA2-CA3*	0.00036806
		SLC39A11	0.00046393
		SLC30A10	0.00063683
	Yoruba	*SCAMP5*	4.17E-05
		SLC39A3	0.00027082
		SLC30A8	0.00044064
	Mandenka	*MTF1*	1.04E-05
		SLC39A3	0.0005911
		MTF2	0.00078435
		SLC30A6	0.00082188
	Mbuti	*MT1F-MT1G-MT1H*	0.00032824
		GPR39	0.00056904
Middle-East	Bedouin	*SLC39A11*	0.00072923
		SLC30A7	0.00076938
	Druze	*SLC39A11*	0.0002821
		CAR13	0.00070799
		CA3	0.00090627
	Mozabite	*CA1*	7.15E-06
		SLC30A8	0.00019367
		GPR39	0.00056815
	Palestinian	*SLC39A11*	1.48E-05
		STAT3	0.00071861
Europe	Adygei	*SLC30A8*	0.00015855
		SLC30A10	0.00027645
		SLC39A11	0.00074135
	Basque	*SLC30A10*	0.0001034
		SLC39A11	0.00092046
	BergamoItalian-Tuscan	*SLC30A8*	4.48E-05
		SLC39A12	0.00041253
		SLC39A11	0.000771
		MT1A	0.000771
		MT1E	0.000771
	French	*SLC39A11*	9.13E-05
		SLC30A8	0.00072794
	Orcadian	*SLC39A11*	0.00011567
		SLC30A10	0.00021496
		SLC30A9	0.00081243
	Russian	*SLC30A1*	1.32E-05
		SLC39A11	0.00015104
		SLC30A6	0.00037718
	Sardinian	*SLC39A11*	0.00044753
		MT2A	0.00066616
Central-South Asia	Balochi	*SLC39A11*	1.55E-05
	Brahui	*SLC39A11*	0.00061849
		SLC39A3	0.00061849
		CA3	0.00090293
		SLC30A9	0.00092172

		SLC39A12	0.00093274
		SLC39A14	0.00094336
	Burusho	*SLC39A11*	0.00012304
		SLC30A10	0.00039715
	Hazara	*SLC39A11*	0.00076257
	Kalash	*SLC39A10*	1.21E-05
		SLC30A8	0.0005821
		SLC39A11	0.00074226
		SLC30A10	0.00085796
	Makrani	*SLC39A11*	1.40E-06
		SLC39A12	2.40E-05
		SLC30A9	0.00036417
	Pathan	*SLC30A7*	9.46E-05
		SLC39A3	0.00037334
		SLC30A9	0.00065901
	Sindhi	*SLC39A11*	0.00015917
		SLC30A9	0.00031178
	Uygur	*SLC30A9*	0.0002744
East Asia	Dai-Lahu	*GPR39*	0.00065143
	Han	*SLC39A11*	0.0005133
	Japanese	*GPR39*	7.51E-05
		SLC30A8	0.00052042
		SLC39A11	0.00058404
		SLC39A9	0.0007399
	Naxi-Yi	*SLC39A11*	8.43E-05
		SLC39A8	0.00016511
		IL6	0.00037186
		SLC30A10	0.00037186
		SLC39A9	0.00064471
	Northern-Han	*SLC39A11*	3.90E-05
		SLC30A10	0.00065104
		MTF2	0.00065104
	She-Miao-Tujia	*SLC39A11*	2.41E-05
		SLC11A1	0.00033635
		GPR39	0.00086595
	Xibo-Mongolian	*MT1F-MT1G-MT1H*	0.0002743
		MTF2	0.00052523
		SLC39A11	0.00087309
	Yakut	*IL6R*	0.00013508
		SLC30A4	0.00096514
Americas	Surui-Karitiana	*SLC39A10*	0.00014355
		SLC39A8	0.00020085
		SLC39A11	0.00041299
Oceania	Bougainville	*SLC39A11*	0.000375
		SLC30A6	0.000375
	Papuan	*SLC30A9*	0.00028702
		SLC39A9	0.00077394

***Table S4.5**: The Zinc-associated genes within the 0.1% tail, as indicated by the F_{ST} selection values for each population. Ordered by most significant.*

Region	Population	Gene	$Relate\ P-value$
Africa	Bantu-speaking	*SLC30A9*	2.83E-05
		CAR13	8.23E-05
		SLC39A12	0.00011711
		SLC39A11	0.00029278
	Biaka	*SLC39A9*	0.00019231
		SCAMP5	0.00024413
		SLC30A10	0.00043508
		SLC30A5	0.00072876
		SLC39A4	0.00086389
		SLC39A5	0.0008651
	Mandenka	*SLC39A12*	4.04E-05
		SLC39A13	0.0002323
		STAT3	0.00056251
		SLC39A2	0.000744
	Mbuti	*GPR39*	9.97E-05
		SLC39A4	0.00015312
		SLC39A13	0.00018123
		SLC39A11	0.00050677
	San	*SLC39A4*	9.96E-05
		SLC39A11	0.00010416
		SLC39A2	0.00021486
		SLC11A1	0.00024216
Middle-East	Bedouin	*GPR39*	5.59E-05
		SLC39A14	0.00023457
		SLC39A4	0.00030943
		SLC39A11	0.00033089
		CA1	0.00048109
	Druze	*SLC39A4*	1.02E-05
		GPR39	3.23E-05
		SLC39A14	0.00041768
		SLC39A11	0.00071243
		CA1	0.0007683
	Mozabite	*CA3*	0.0002021
		SLC39A4	0.00020628
		SLC39A11	0.00028293
		CA1	0.00028765
		GPR39	0.00036971
	Palestinian	*SLC39A4*	2.85E-05
		GPR39	3.72E-05
		SLC39A14	5.81E-05
		CA1	0.0004049
		SLC39A11	0.00064181
		SCAMP5	0.00080733
Europe	Adygei	*SLC39A4*	2.30E-05
		SLC30A9	0.00011976
		SLC39A8	0.0001604
		GPR39	0.00016129
		SLC39A11	0.00034286
		SLC30A2	0.00060677
	Basque	*SLC39A4*	1.41E-05
		GPR39	0.00013401
		SLC39A11	0.00032407
		SLC39A14	0.00099308

	BergamoItalian-Tuscan	*SLC39A4*	8.84E-06
		GPR39	8.94E-05
		SLC30A9	0.00030164
		SLC39A14	0.00052329
		SLC30A8	0.00064347
		SLC39A11	0.00068438
	French	*SLC39A4*	5.58E-06
		GPR39	5.35E-05
		SLC39A14	0.00051554
	Orcadian	*SLC39A4*	3.19E-05
		SLC30A10	9.97E-05
		GPR39	0.00028039
		SLC39A11	0.00047506
		SLC30A9	0.00080692
		SCAMP5	0.00080692
	Russian	*SLC39A4*	6.43E-06
		GPR39	6.47E-05
		SLC30A10	0.0001957
		SLC39A8	0.00033289
		SLC39A14	0.00037319
		MTF2	0.00058517
		SLC39A11	0.00073398
	Sardinian	*SLC39A4*	1.33E-05
		GPR39	5.57E-05
		SLC39A14	0.0009194
Central-South Asia	Balochi	*SLC39A4*	2.89E-05
		GPR39	0.00013587
		SLC30A2	0.00017753
		SLC39A11	0.00022464
		SLC39A14	0.00032267
		SLC30A9	0.00066326
	Brahui	*SLC39A4*	7.98E-06
		GPR39	2.63E-05
		SLC39A11	0.00045037
		SLC39A14	0.00047577
		SLC30A9	0.00084037
		SLC11A1	0.00085518
	Burusho	*SLC39A4*	8.76E-06
		SLC30A9	5.38E-05
		GPR39	5.78E-05
		SLC39A11	0.00020162
		MTF2	0.00035538
		SLC30A8	0.00039795
		SLC39A14	0.00053656
		SLC30A2	0.0009226
	Hazara	*SLC39A4*	7.24E-06
		SLC30A9	2.23E-05
		GPR39	9.24E-05
		MTF2	9.77E-05
		SLC39A14	0.00085186
	Kalash	*SLC39A4*	1.72E-05
		SLC39A11	9.98E-05
		GPR39	0.00013799
		MTF2	0.0002331
		SLC30A9	0.00033323
		SLC39A14	0.00044862
		SLC30A10	0.00085675
	Makrani	*SLC39A4*	3.95E-06
		SLC11A1	0.0002031

		GPR39	0.00021093
		SLC39A11	0.00033045
		SLC30A9	0.00055087
		SLC39A14	0.00071355
	Pathan	*SLC39A4*	5.47E-06
		GPR39	3.47E-05
		SLC30A9	6.08E-05
		MTF2	0.00028829
		SLC39A11	0.00040321
		SLC39A14	0.00044799
		SLC30A2	0.00066599
		SLC11A1	0.00076262
	Sindhi	*SLC39A4*	1.03E-05
		SLC39A11	2.99E-05
		GPR39	5.83E-05
		SLC30A2	0.00027912
		SLC30A9	0.00050259
East Asia	Dai-Lahu	*SLC30A9*	2.12E-05
		SLC39A4	7.82E-05
		GPR39	0.00047748
		SLC30A3	0.00073958
		SLC39A8	0.00089389
	Han	*SLC30A9*	3.55E-06
		SLC39A4	4.34E-05
		GPR39	0.00020004
		MTF2	0.00075658
	Japanese	*SLC39A4*	6.69E-05
		SLC30A9	0.00014582
		GPR39	0.00044395
		MTF2	0.00071933
		SLC39A8	0.00077016
	Oroqen-Hezhen-Daur	*SLC30A9*	1.51E-05
		SLC39A4	3.99E-05
		SLC30A2	0.00039712
		GPR39	0.00056561
	Naxi-Yi	*SLC39A4*	9.03E-05
		SLC30A9	0.00010709
		SLC39A8	0.00031326
		GPR39	0.00042359
		MTF2	0.00052899
		SLC30A3	0.00065973
	NorthernHan-Tu	*SLC39A4*	5.98E-05
		SLC30A9	7.01E-05
		SLC39A8	0.00035757
		GPR39	0.00056756
		IL6R	0.00092879
		SLC30A3	0.00098103
	She-Miao-Tujia	*SLC30A9*	2.05E-05
		SLC39A4	4.75E-05
		GPR39	0.00034618
		MTF2	0.0006867
	Xibo-Mongolian	*SLC30A9*	2.66E-05
		SLC39A4	5.49E-05
		GPR39	0.0002492
		SLC30A3	0.0004534
		IL6R	0.00064751
	Yakut	*SLC30A9*	1.71E-05
		SLC39A4	3.57E-05
		SLC30A2	0.0003788

		GPR39	0.00091171
Americas	Maya	*SLC30A9*	8.30E-05
		STAT3	0.00028013
		SLC39A11	0.00038942
		SLC30A2	0.00043729
		GPR39	0.00049596
		SLC39A4	0.00074664
	Pima	*STAT3*	0.00010471
		SLC39A11	0.00010471
		SLC30A2	0.00010471
		SLC39A4	0.00083811
	Surui-Karitiana	*SLC30A9*	0.00012002
		SLC30A2	0.00030537
		SLC39A9	0.00050277
		SLC39A11	0.00052132
		SLC39A4	0.00077086
		SLC39A8	0.0008551
		STAT3	0.0009046
Oceania	Bougainville	*SLC30A9*	8.23E-05
		GPR39	0.00034338
		SLC39A11	0.00040821
		SLC39A4	0.00043659
	Papuan	*SLC39A11*	9.64E-05
		SLC30A9	0.00025195
		SLC39A4	0.00032253

***Table S4.6**: The Calcium-associated genes within the 0.1% tail, as indicated by the **Relate** selection values for each population. Ordered by most significant.*

Region	**Population**	**Gene**	$\boldsymbol{Relate\ P-value}$
Africa	Bantu-speaking	*SLC8A2*	0.00011146
		SLC8A1	0.00023012
		CYP24A1	0.00044936
		ATP2B1	0.00080919
	Biaka	*DGKD*	1.97E-05
		SLC34A3	0.00027589
		ATP2B2	0.00030009
		TRPM2	0.00050177
		SLC8A1	0.0005978
	Mandenka	*VWA8*	0.00011987
		ATP2B2	0.00018742
		TRPM2	0.00026067
		SLC8A1	0.00063471
	Mbuti	*ATP2B2*	0.00048462
		RGS2	0.00048462
Middle-East	Bedouin	*ATP2B2*	0.00022835
		ATP2B4	0.00053567
		SLC8A1	0.00086055
	Druze	*SLC8A1*	0.00025185
		ATP2B4	0.00030962
	Mozabite	*ATP2B2*	1.38E-05
		VWA8	0.0001365
		ATP2B4	0.00019367
		CASR	0.00071448
	Palestinian	*ATP2B2*	0.00017141
		TRPM2	0.00071861

		SLC8A1	0.00073645
		ATP2B3	0.00080695
Europe	Adygei	*SLC8A1*	0.00057084
		GATA3	0.00074135
	Basque	*ATP2B2*	1.35E-04
	BergamoItalian-Tuscan	*SLC8A1*	4.48E-05
		ATP2B4	0.00041253
		TRPM2	0.00049621
		ATP2B2	0.00054845
	French	*ATP2B2*	7.51E-05
		ATP2B4	0.00050016
		SLC8A1	0.00072794
		TRPM2	0.00072794
		ATP2B1	0.00072794
	Orcadian	*CYP24A1*	0.00011567
		ATP2B2	0.00059697
		DGKD	0.00081243
	Russian	*ATP2B2*	0.00027103
		DGKD	0.00084244
		SLC8A1	0.00088073
	Sardinian	*ATP2B2*	2.10E-07
		SLC8A1	0.00012589
		VWA8	0.0002822
		ATP2B4	0.00065672
Central-South Asia	Balochi	*VWA8*	0.00085698
		SLC8A1	0.00091688
	Brahui	*SLC8A2*	0.00032201
	Burusho	*ATP2B2*	9.35E-05
		SLC8A3	0.00038816
	Hazara	*SLC8A1*	0.00067934
	Kalash	*SLC12A3*	2.21E-05
		DGKD	0.0005821
		SLC8A1	0.00085796
		SLC8A3	0.00096965
		ATP2B2	0.00096965
	Makrani	*SLC8A1*	2.40E-05
		ATP2B2	4.97E-05
	Pathan	*ATP2B2*	2.08E-05
		SLC8A1	0.00025819
		SLC8A3	0.0004843
	Sindhi	*VWA8*	0.00036757
		ATP2B4	0.00089124
	Uygur	*ATP2B2*	5.23E-05
		SLC8A1	0.0002744
East Asia	Dai-Lahu	*SLC8A3*	4.43E-05
		SLC8A1	7.39E-05
	Han	*TRPM2*	0.00017549
		ATP2B2	0.00065698
	Japanese	*DGKD*	0.00020117
		ATP2B2	0.0002387
		SLC8A3	0.00049776
		TRPM2	0.00053484
		SLC8A1	0.0007534
	Oroqen-Hezhen-Daur	*TRPM2*	0.00075899
	Naxi-Yi	*SLC8A3*	0.00037186
		VWA8	0.00042524
		ATP2B2	0.0004816
		SLC8A1	0.00066865
	NorthernHan-Tu	*SLC8A3*	3.90E-05

	She-Miao-Tujia	*TRPM2*	0.00021958
		GATA3	0.00024253
		ATP2B4	0.00024253
		SLC8A3	0.00072692
		ATP2B2	0.00083491
		SLC8A1	0.00086595
	Xibo-Mongolian	*SLC8A3*	6.74E-05
		ATP2B2	0.00052523
		GATA3	0.00087309
	Yakut	*SLC8A1*	0.000452
Americas	Maya	*SLC8A1*	4.17E-05
		ATP2B1	0.00015384
	Pima	*ATP2B2*	9.06E-06
	Surui-Karitiana	*SLC8A1*	0.00065956
Oceania	Bougainville	*ATP2B2*	0.000375
	Papuan	*SLC8A1*	1.26E-05
		ATP2B2	1.97E-05

***Table S4.7**: The Calcium-associated genes within the 0.1% tail, as indicated by the F_{ST} selection values for each population.* *Ordered by most significant.*

Region	Population	Gene	F_{ST} $P-value$
Africa	Bantu-speaking	*KCNJ10*	1.38E-05
		ATP2B4	0.00017082
		CYP24A1	0.00084385
		ATP2B2	0.00085097
		SLC34A3	0.00085097
		SLC8A1	0.000922
	Biaka	*SLC8A1*	3.05E-05
		ATP2B4	8.39E-05
		CYP24A1	0.00044008
		GCKR	0.00050054
		GATA3	0.00051639
		RGS2	0.00087852
		ATP2B2	0.00095405
		TRPM2	0.00098545
	Mandenka	*ATP2B2*	7.75E-08
		SLC8A1	3.00E-05
		KCNJ10	3.01E-05
		GCKR	0.00032727
	Mbuti	*ATP2B4*	0.00099714
	San	*TRPM2*	0.00024216
		SLC8A1	0.00029778
		ATP2B2	0.00080396
Middle-East	Bedouin	*ATP2B2*	8.84E-05
		ATP2B4	0.00020166
		CYP24A1	0.00071441
		SLC8A2	0.00086433
	Druze	*ATP2B2*	0.00035581
		SLC8A2	0.00078781
	Mozabite	*ATP2B4*	8.32E-05
		ATP2B2	9.73E-05
	Palestinian	*ATP2B2*	0.0003236
		ATP2B4	0.00038601
		CYP24A1	0.00094864
Europe	Adygei	*ATP2B4*	0.00038826

		ATP2B2	0.00057071
		SLC8A2	0.00077579
	Basque	*ATP2B2*	0.00026722
		ATP2B4	0.00068692
		SLC8A2	0.00086331
	BergamoItalian-Tuscan	*ATP2B4*	0.0002526
		ATP2B2	0.00037792
		SLC8A2	0.00050789
	French	*ATP2B2*	7.13E-05
		SLC8A1	0.0002476
		ATP2B4	0.00031481
		SLC8A2	0.00073577
	Orcadian	*ATP2B2*	0.00035948
		SLC8A2	0.00060977
	Russian	*ATP2B2*	0.0001368
		ATP2B4	0.00018927
	Sardinian	*SLC8A2*	0.00014296
		SLC8A1	0.00050004
		ATP2B4	0.00073068
		ATP2B2	0.00096944
Central-South Asia	Balochi	*ATP2B4*	0.0001541
		ATP2B2	0.00035706
	Brahui	*ATP2B2*	0.00067809
		SLC8A1	0.00099182
	Burusho	*ATP2B4*	0.00048762
		ATP2B2	0.0009027
		SLC8A2	0.00093919
	Kalash	*SLC8A2*	0.00083374
	Makrani	*SLC8A1*	0.00065854
		ATP2B4	0.00086095
		ATP2B2	0.00087364
	Pathan	*ATP2B4*	0.00039607
		SLC8A2	0.00041262
		ATP2B2	0.00081051
	Sindhi	*SLC8A1*	0.0003684
		ATP2B2	0.00078334
		SLC8A2	0.00082772
	Uygur	*ATP2B4*	0.00029722
		SLC8A1	0.00034663
		SLC8A2	0.00038054
		SLC8A3	0.00067718
East Asia	Han	*CYP24A1*	0.00070583
		ATP2B2	0.00090958
	Japanese	*CASR*	0.00076711
	Oroqen-Hezhen-Daur	*CYP24A1*	0.00082205
	Naxi-Yi	*CASR*	0.0007695
	NorthernHan-Tu	*ATP2B2*	0.00060215
	Yakut	*ATP2B2*	0.0008125
Americas	Maya	*SLC8A1*	0.0008765
	Pima	*CASR*	0.0018747
	Surui-Karitiana	*SLC8A1*	0.0008551
Oceania	Papuan	*ATP2B2*	0.00060372
		SLC8A3	0.00081085

***Table S4.8**: The Selenium-associated genes within the 0.1% tail, as indicated by the **Relate selection values for each population**. Ordered by most significant.*

Region	Population	Gene	$Relate\ P-value$
Africa	Bantu-speaking	*LRP8*	8.80E-05
		PRKG1	0.00011146
		EEFSEC	0.00016167
		AKAP6	0.00019916
		SCLY	0.00066041
		SGCD	0.00069572
		SELENOP	0.00080919
	Biaka	*SGCD*	8.77E-05
		SELENOS	0.00011324
		LHFPL2	0.00014782
		AKAP6	0.00027589
		KCNMA1	0.00032071
		PRKG1	0.0005978
	Yoruba	*SELENOM*	5.87E-06
		PRKG1	0.00016458
		LRP8	0.00072398
		AKAP6	0.00078661
		KCNMA1	0.00078661
		SELENOS	0.00082895
	Mandenka	*LRP8*	1.04E-05
		LHFPL2	0.00011467
		SECISBP2	0.00027311
		LRP2	0.00034801
		KCNMA1	0.00053903
		AKAP6	0.0005911
		SELENOP	0.00071216
		PRKG1	0.00094662
	Mbuti	*SGCD*	8.72E-05
		KCNMA1	0.00020261
		EEFSEC	0.00020261
		TXNRD1	0.00032824
		SELENOI	0.00077346
		LRP2	0.00089587
Middle-East	Bedouin	*PRKG1*	0.00022585
		AKAP6	0.00060058
		SGCD	0.00062467
	Druze	*AKAP6*	0.00037687
		SGCD	0.00069323
	Mozabite	*AKAP6*	4.16E-05
		SCLY	9.09E-05
		GPx7	0.00028962
		TXNRD3	0.00057637
		SGCD	0.00062854
	Palestinian	*PRKG1*	6.83E-05
		SGCD	9.50E-05
		SELENOF	0.00010856
		AKAP6	0.00019775
		SELENOS	0.00071861
Europe	Adygei	*AKAP6*	4.62E-05
		KCNMA1	0.0001784
		PRKG1	0.00058113
		EEFSEC	0.00065646
		TXNRD3	0.00078344

	Basque	*TXNRD3*	4.79E-05
		EEFSEC	5.59E-05
		AKAP6	0.00035174
		SELENOF	0.00040146
		SGCD	0.00040146
		PRKG1	0.00047061
	BergamoItalian-Tuscan	*KCNMA1*	0.00018138
		SGCD	0.00067804
		AKAP6	0.000771
		GPx2	0.0008142
	French	*SELENOF*	0.00097709
	Orcadian	*SELENOF*	0.00021496
		SEPSECS	0.00069701
	Russian	*KCNMA1*	0.000137
		SELENOF	0.00015104
		SEPSECS	0.00015828
		SCLY	0.00026209
		SECISBP2	0.00040399
		LRP2	0.00086052
		TXNRD1	0.00087605
	Sardinian	*SECISBP2*	3.27E-05
		PRKG1	0.0003378
		SELENOF	0.00047208
		AKAP6	0.00052747
		SGCD	0.00065672
Central-South Asia	Balochi	*PRKG1*	0.000145
		LRP8	0.00027678
		SELENOS	0.00027678
		GPx2	0.00029206
		EEFSEC	0.00035898
		ELAVL1	0.00085698
	Brahui	*GPx2*	0.00010148
		LRP2	0.00025177
		SELENOF	0.00076429
	Burusho	*AKR7L*	9.83E-06
		PRKG1	0.00012555
		ELAVL1	0.00023813
		AKAP6	0.00051078
		ARSB	0.00064344
		SGCD	0.00069345
	Hazara	*PRKG1*	5.56E-05
		ARSB	0.00017433
		SELENOP	0.00017433
		TXNRD1	0.00017433
		LRP2	0.00034417
		SGCD	0.0004472
	Kalash	*PRKG1*	0.00014809
	Makrani	*GPx2*	9.61E-06
		DIO2	0.00028549
	Pathan	*SELENOP*	4.09E-05
		SELENOF	0.00012857
		LRP2	0.00013172
		PRKG1	0.00025819
	Sindhi	*KCNMA1*	0.00036757
		SGCD	0.00063874
		SELENOF	0.00089124
		PRKG1	0.00089124
East Asia	Dai-Lahu	*PRKG1*	0.00025059
		SGCD	0.00025059

		KCNMA1	0.0008572
		TXNRD2	0.00087044
	Han	*PRKG1*	6.54E-05
		KCNMA1	0.00023142
		SGCD	0.00033036
	Japanese	*LHFPL2*	0.00015814
		KCNMA1	0.00033413
		PRKG1	0.00049776
		ARSB	0.00049776
		SELENOI	0.00052042
	Oroqen-Hezhen-Daur	*PRKG1*	0.00023563
		ARSB	0.00023563
		AKAP6	0.00023563
		SGCD	0.00036709
	Naxi-Yi	*PRKG1*	2.17E-05
		SELENOI	0.00037186
		AKAP6	0.00041152
		KCNMA1	0.00047387
		SGCD	0.00071255
	NorthernHan-Tu	*KCNMA1*	3.90E-05
		PRKG1	0.00015139
		SELENOI	0.0003887
		AKAP6	0.0003887
		SELENOS	0.0003887
		SGCD	0.00079954
	She-Miao-Tujia	*PRKG1*	0.00013476
		KCNMA1	0.00024725
		SGCD	0.000598
	Xibo-Mongolian	*KCNMA1*	0.00016803
		PRKG1	0.0002743
		SELENOP	0.0002743
		LRP8	0.00052523
		ARSB	0.00052523
		AKAP6	0.00087309
	Yakut	*KCNMA1*	1.46E-05
		AKAP6	7.62E-05
		GPx7	0.00013104
		DIO2	0.00050707
		SEPHS1	0.00058843
		SGCD	0.00072576
		PRKG1	0.00096514
Americas	Maya	*LHFPL2*	0.00075386
	Pima	*TRNAU1AP*	0.00018445
		SELENOI	0.00077597
	Surui-Karitiana	*AKAP6*	3.96E-05
Oceania	Bougainville	*SGCD*	0.00021637
		DMGDH	0.00039673
		BHMT	0.00039673
	Papuan	*DIO2*	0.00027674
		AKAP6	0.00063515

***Table S4.9**: The Selenium-associated genes within the 0.1% tail, as indicated by the F_{ST} selection values for each population.* *Ordered by most significant*.

Region	Population	Gene	F_{ST} $P-value$
Africa	Bantu-speaking	*LHFPL2*	4.99E-06
		PRKG1	2.11E-05
		EEFSEC	2.53E-05
		SGCD	0.00017082
		GPx3	0.00018601
		KCNMA1	0.00025396
		TXNRD2	0.0004741
		ARSB	0.00050132
		SELENOI	0.00080305
		TRU-TCA1-1	0.00095671
	Biaka	*LHFPL2*	8.33E-05
		EEFSEC	0.00010422
		SELENOI	0.00011036
		SGCD	0.00045229
		LRP8	0.00066915
		SEPHS2	0.00071412
		KCNMA1	0.00094741
	Mandenka	*TXNRD2*	9.72E-05
		SARS2	0.00015415
		SGCD	0.00040891
		RPL30	0.00054919
		EIF4A3	0.00063262
		ARSB	0.00070512
	Mbuti	*TRU-TCA2-1*	6.63E-05
		LHFPL2	7.44E-05
		SELENOH	0.000126
		ARSB	0.0002288
		CELF1	0.00030309
		SEPHS1	0.00052952
		SELENOI	0.00062595
		TRU-TCA3-1	0.00062595
	San	*LRP8*	2.15E-05
		LHFPL2	3.08E-05
		SGCD	0.00010027
		GPx3	0.00037697
		PSTK	0.00037697
		AKAP6	0.00077386
		JMY	0.00077386
Middle-East	Bedouin	*EEFSEC*	8.33E-05
		AKAP6	0.00025165
		LHFPL2	0.0004813
		SGCD	0.00050928
		LRP8	0.00056435
		PRKG1	0.00066552
		JMY	0.00066587
		GPx4	0.00070823
		TXNRD2	0.00089399
	Druze	*EEFSEC*	0.00016841
		SGCD	0.00017835
		AKAP6	0.00071065
		LHFPL2	0.0007289
		SELENOS	0.00076221
		ARSB	0.00090391

		TXNRD2	0.00091089
	Mozabite	*EEFSEC*	1.29E-05
		TXNRD2	0.00026529
		AKAP6	0.00065171
		SGCD	0.00074499
	Palestinian	*JMY*	0.00013105
		LHFPL2	0.00021016
		AKAP6	0.00025078
		DIO1	0.0003277
		DIO2	0.00046738
		PRKG1	0.00079409
		ARSB	0.00087271
		TXNRD2	0.00088318
Europe	Adygei	*EEFSEC*	0.00010962
		SELENOS	0.00038826
		AKAP6	0.00062343
		LRP8	0.0009683
	Basque	*SELENOS*	0.00061776
		AKAP6	0.00061776
		LRP8	0.0006782
		EEFSEC	0.00068692
		DIO1	0.00070659
		PRKG1	0.00095054
		SGCD	0.00096781
	BergamoItalian-Tuscan	*SGCD*	1.07E-05
		EEFSEC	0.0001355
		AKAP6	0.00036307
		DIO2	0.00068438
		PRKG1	0.00077155
		LRP8	0.00081451
	French	*SGCD*	0.00023628
		SELENOS	0.00044557
		TXNRD3	0.00054689
		DIO2	0.00063582
		DIO1	0.0006635
		LRP8	0.00086606
		EEFSEC	0.00087608
	Orcadian	*ARSB*	7.92E-05
		EEFSEC	0.00017297
		GPx2	0.0002503
		SGCD	0.0002571
		LRP8	0.00038381
		AKAP6	0.00040702
		GPx3	0.00060977
	Russian	*SELENOS*	4.01E-05
		LRP8	0.00012788
		EEFSEC	0.00021035
		AKAP6	0.0002722
		SGCD	0.00033211
		PRKG1	0.00053471
		DMGDH	0.00076172
		DIO1	0.00092937
	Sardinian	*EEFSEC*	0.00013033
		AKAP6	0.00018068
		KCNMA1	0.00060425
		LRP2	0.00075641
		SCLY	0.00097505
	Balochi	*SELENOS*	0.00015839
		EEFSEC	0.00032175

		SGCD	0.00044376
		PRKG1	0.0004747
		SELENON	0.00088683
	Brahui	*SELENOS*	8.04E-05
		AKAP6	0.0001368
		EEFSEC	0.00033199
		SGCD	0.0004361
		DIO2	0.00064879
		PRKG1	0.00067809
	Burusho	*DIO1*	0.00013009
		AKAP6	0.00020316
		DIO2	0.00039795
		EEFSEC	0.00042983
		SELENOS	0.00066096
	Hazara	*SELENOS*	1.86E-05
		EEFSEC	0.000255
		DIO1	0.00041576
		PRKG1	0.00054207
		SELENOI	0.00060433
		ARSB	0.0008675
	Kalash	*EEFSEC*	0.00013799
		SELENOS	0.0002105
		DIO1	0.00032655
		LRP8	0.00047725
		DIO2	0.00058909
		ARSB	0.00097016
		DMGDH	0.00097016
	Makrani	*SGCD*	1.18E-05
		SELENOS	0.00013145
		AKAP6	0.00022042
		PRKG1	0.0002227
		GPx2	0.00025279
		EEFSEC	0.00029101
		DIO2	0.00042861
	Pathan	*DIO1*	0.00027736
		AKAP6	0.00028829
		SGCD	0.00054401
		EEFSEC	0.00056375
		SELENOS	0.00061362
		PRKG1	0.00069817
		SEPHS1	0.00079032
	Sindhi	*SELENOS*	0.00013278
		EEFSEC	0.00029401
		AKAP6	0.00031698
		DIO1	0.00046362
		SELENON	0.00047021
		PRKG1	0.0007474
		SEPHS1	0.00091271
	Uygur	*PRKG1*	0.00019889
		SEPHS1	0.00025021
		EEFSEC	0.00029722
		LRP8	0.00044421
		DIO1	0.00057239
		SELENOS	0.00062677
		DIO2	0.00062677
		SGCD	0.00077625
		ARSB	0.00091985
East Asia	Dai-Lahu	*SEPHS1*	0.00013178
		PRKG1	0.00041625

	SELENOI	0.00045413
	SEPHS2	0.00054729
	DMGDH	0.00089389
	BHMT2	0.00089389
Han	*DIO1*	0.00024443
	SELENOI	0.00030395
	EEFSEC	0.00041875
	PRKG1	0.00045551
	TXNRD2	0.00067604
	SEPHS2	0.00068557
	SELENOW	0.0007888
	SELENON	0.00080415
Japanese	*SELENOW*	6.91E-05
	SEPHS2	0.00023089
	PRKG1	0.00037851
	DIO1	0.00047758
	SELENOI	0.00051479
	EEFSEC	0.00055858
	KCNMA1	0.00096672
Oroqen-Hezhen-Daur	*KCNMA1*	0.00015128
	SELENOI	0.00029749
	SELENOW	0.00035763
	DIO1	0.00041683
	SEPHS2	0.00047206
	PRKG1	0.00066189
	SELENOS	0.00085593
Naxi-Yi	*SEPHS2*	0.00010969
	PRKG1	0.0001864
	SELENOW	0.00036851
	SEPHS1	0.00048678
	SELENOI	0.00057711
	SELENOS	0.0008885
NorthernHan-Tu	*PRKG1*	0.00011442
	SEPHS2	0.00020983
	DIO1	0.00035868
	SELENOS	0.00042754
	SEPHS1	0.00060159
	EEFSEC	0.00073565
	SELENOI	0.00098636
	KCNMA1	0.00098636
She-Miao-Tujia	*PRKG1*	2.05E-05
	SELENOI	0.00037182
	EEFSEC	0.00052419
	GPx1	0.00056207
	KCNMA1	0.00090569
Xibo-Mongolian	*PRKG1*	1.00E-05
	SEPHS2	2.66E-05
	SELENOI	0.00042129
	SELENOW	0.00048511
	SEPHS1	0.00058305
	DMGDH	0.00060539
	EEFSEC	0.00079502
	JMY	0.00087862
Yakut	*PRKG1*	0.00010185
	DIO1	0.0002481
	SEPHS2	0.00031133
	EEFSEC	0.00049965
	AKAP6	0.00058379
	SELENOS	0.00095554

		KCNMA1	0.00099665
Americas	Maya	*GPx3*	0.00014139
		SGCD	0.00031806
		DIO1	0.00062126
		DIO2	0.00067055
		SEPHS2	0.00083053
	Pima	*GPx3*	0.00010471
		SELENON	0.00010471
		SGCD	0.00044989
		SEPHS2	0.00044989
		PRKG1	0.00066221
		DIO2	0.00086237
	Surui-Karitiana	*GPx3*	0.00012002
		SGCD	0.00012002
		SEPHS2	0.00027794
		LHFPL2	0.00030537
		AKAP6	0.00052132
		SELENON	0.00059791
		SELENOM	0.00090882
Oceania	Bougainville	*AKAP6*	0.00017199
		SGCD	0.00034338
		DIO2	0.00034338
		LRP8	0.00034338
		TRU-TCA2-1	0.00062923
	Papuan	*SGCD*	5.36E-05
		PRKG1	0.00070944
		KCNMA1	0.00081085

***Table S4.10**: The Iron-associated genes within the 0.1% tail, as indicated by the **Relate** selection values for each population. Ordered by most significant.*

Region	**Population**	***Gene***	***Relate P − value***
Africa	Biaka	*TMPRSS6*	7.18E-05
		ACO1	0.00030009
		ARHGEF3	0.00046393
		MYB	0.00046393
		HIF1A	0.00048144
	Yoruba	*HIF1A*	0.00035819
		SLC48A1	0.00063471
		PANK2	0.00082895
	Mandenka	*HAMP*	0.00012474
		FTMT	0.00015476
	Mbuti	*ACO2*	0.00032824
		PLA2G6	0.00077346
Middle-East	Bedouin	*EPAS1*	0.00013935
		HBS1L	0.0002138
	Druze	*WDR45*	0.00010661
		CP	0.0002821
		FTMT	0.00058571
		FA2H	0.00058571
		TMPRSS6	0.00058571
		SLC48A1	0.00080002
	Mozabite	*HIF1A*	0.00078773
Europe	Adygei	*C19orf12*	2.01E-05
		CFAP251	9.28E-05
		ARHGEF3	0.00015855

		FTMT	0.00099437
	Basque	*HIF1A*	2.43E-06
		ARHGEF3	2.47E-05
		MYB	0.00085462
	BergamoItalian-Tuscan	*HIF1A*	0.00041253
		LTF	0.00046929
		SLC40A1	0.00049621
	French	*HIF1A*	0.00026554
		SLC11A2	0.00085614
	Orcadian	*FTMT*	0.00012535
		C19orf12	0.00021496
		EPAS1	0.00059697
	Russian	*FTMT*	0.00084244
	Sardinian	*FA2H*	0.00052747
Central-South Asia	Balochi	*EPAS1*	0.00027678
		FTMT	0.00091688
	Brahui	*FTMT*	1.92E-05
		PANK2	0.00061849
		EPAS1	0.00094336
	Burusho	*HIF1A*	0.0002737
		PANK2	0.00064344
	Hazara	*ARHGEF3*	0.00046092
	Kalash	*CFAP251*	0.0005821
	Makrani	*FTMT*	0.00050511
	Pathan	*HIF1A*	0.00029385
	Sindhi	*HIF1A*	2.28E-05
	Uygur	*SLC40A1*	1.62E-05
		PLA2G6	7.10E-05
East Asia	Dai-Lahu	*ARHGEF3*	6.80E-05
		LCN2	0.00051035
	Han	*LCN2*	0.00047733
		RHOA	0.00047733
		TAOK1	0.00051001
	Japanese	*SLC40A1*	0.0001792
	Oroqen-Hezhen-Daur	*SLC40A1*	0.00023563
		ARHGEF3	0.00023563
	NorthernHan-Tu	*RHOA*	0.0003887
		FTMT	0.00065104
		TFRC	0.00074298
	She-Miao-Tujia	*FTMT*	0.00021508
		SLC11A1	0.00033635
		TFRC	0.00072692
	Xibo-Mongolian	*RHOA*	0.00052523
		FTMT	0.0009191
		SLC17A1	0.00098995
	Yakut	*FTMT*	3.37E-06
		SLC40A1	0.00042053
		PLA2G6	0.00042053
		ACO1	0.00096514
Americas	Maya	*ARHGEF3*	0.00028516
		SLC48A1	0.00036477
		TF	0.00047791
	Pima	*SLC17A1*	0.00071578
Oceania	Papuan	*HIF1A*	0.00027674
		HBS1L	0.00027674

***Table S4.11**: The Iron-associated genes within the 0.1% tail, as indicated by the F_{ST} selection values for each population.* *Ordered by most significant.*

Region	Population	Gene	F_{ST} $P-value$
Africa	Bantu-speaking	*TAOK1*	3.42E-05
		ATP13A2	0.00019915
		HJV	0.00025213
		CFAP251	0.00031076
		TFRC	0.00048283
		ARHGEF3	0.00081567
	Biaka	*ARHGEF3*	0.00050083
		FTL	0.00074618
	Mandenka	*FTL*	1.99E-05
		HJV	2.91E-05
		LCN2	0.0001794
		TMPRSS6	0.00050372
		ACO1	0.00059993
		TF	0.00097832
	Mbuti	*SLC11A2*	0.00037127
		LTF	0.00051764
	San	*ACO2*	0.00021486
		SLC11A1	0.00024216
		ARHGEF3	0.00030976
		ACO1	0.00077386
Middle-East	Bedouin	*ARHGEF3*	3.43E-05
		LCN2	0.00075733
		LTF	0.00079045
	Druze	*ARHGEF3*	0.00035974
		EPAS1	0.00045537
		FTH1	0.00090651
		LCN2	0.00091089
	Mozabite	*ARHGEF3*	0.00044427
		MYB	0.00057807
	Palestinian	*ARHGEF3*	2.55E-05
		LCN2	0.00044637
		EPAS1	0.0005172
		HIF1A	0.00084412
		CFAP251	0.00084851
Europe	Adygei	*FTMT*	7.54E-05
		TMPRSS6	0.00023855
		ARHGEF3	0.00027147
		EPAS1	0.00080734
	Basque	*ARHGEF3*	0.00016136
		HIF1A	0.00024923
		FTMT	0.00040115
		FTH1	0.00078535
	BergamoItalian-Tuscan	*LTF*	0.00018177
		ARHGEF3	0.00058054
		HIF1A	0.00059546
		FTMT	0.00087207
	French	*CFAP251*	0.00013879
		ARHGEF3	0.0002749
		EPAS1	0.0008926
		HIF1A	0.00095844
	Orcadian	*SLC40A1*	0.0005459
		ARHGEF3	0.00067044
		EPAS1	0.00080548

	Russian	*SLC40A1*	0.00044697
	Sardinian	*FTMT*	0.00063005
		FTH1	0.00095923
Central-South Asia	Balochi	*ARHGEF3*	2.34E-05
		FTMT	0.00032236
		HIF1A	0.00060766
	Brahui	*ARHGEF3*	6.51E-05
		FTMT	0.00031265
		SLC11A1	0.00085518
	Burusho	*ARHGEF3*	0.00029444
	Hazara	*ARHGEF3*	0.00012885
	Kalash	*TMPRSS6*	0.00027137
		LTF	0.00027137
		HFE	0.00029735
		ARHGEF3	0.00033265
	Makrani	*ARHGEF3*	8.20E-05
		SLC11A1	0.0002031
		FTMT	0.00075807
	Pathan	*SLC11A1*	0.00076262
		FTMT	0.00098729
	Sindhi	*TMPRSS6*	0.00039545
		EPAS1	0.00079401
	Uygur	*SLC11A1*	0.00023123
		LTF	0.00038054
		SLC40A1	0.00064543
		HBS1L	0.00067718
East Asia	Dai-Lahu	*FTMT*	4.85E-05
		RHOA	0.00066748
		EPAS1	0.00076375
	Han	*RHOA*	8.62E-05
		FTMT	0.00080415
		STEAP3	0.00082102
	Japanese	*RHOA*	6.91E-05
	Naxi-Yi	*SLC40A1*	0.00019231
		RHOA	0.00038495
		CFAP251	0.00078732
		SLC17A1	0.00080911
		FECH	0.00098344
	NorthernHan-Tu	*RHOA*	9.99E-05
	She-Miao-Tujia	*RHOA*	1.38E-05
		ARHGEF3	0.00040273
		TFRC	0.00079686
		SLC40A1	0.00079686
	Xibo-Mongolian		
		RHOA	0.00015719
		SLC40A1	0.00101931
	Yakut	*CFAP251*	0.00071529
		FTMT	0.00082692
		SLC40A1	0.00085024
Americas	Maya	*TMPRSS6*	4.90E-05
		RHOA	0.00020031
		SLC40A1	0.00024802
		TF	0.00099565
	Pima	*RHOA*	0.00010471
		SLC40A1	0.00010471
		SLC48A1	0.00018886
		TMPRSS6	0.00030107
		HIF1A	0.00030813
		EPAS1	0.00085055

	Surui-Karitiana	*ACO2*	0.00086237
		RHOA	0.00023151
		C19orf12	0.00030537
		STEAP3	0.00030537
Oceania	Bougainville	*TFRC*	3.80E-05
		TMPRSS6	8.23E-05
		ACO1	0.00016154
		RHOA	0.00072992
		HIF1A	0.00096615
	Papuan	*ACO1*	5.36E-05
		TMPRSS6	9.64E-05

***Table S4.12**: The Iodine-associated genes within the 0.1% tail, as indicated by the **Relate** selection values for each population. Ordered by most significant.*

Region	Population	Gene	***Relate P − value***
Africa	Bantu-speaking	*TTR*	0.00016167
		THRB	0.00069572
		SLC16A10	0.00029907
	Biaka	*SLCO1C1*	0.00010211
		SULT6B1	0.00014782
	Mandenka	*SECISBP2*	0.00027311
		TTR	0.00034801
		SECISBP2	0.00027311
	Mbuti	*SULT6B1*	0.00032824
Middle-East	Druze	*SULT6B1*	0.00064949
	Mozabite	*TPO*	2.50E-05
		TRIP4	0.00024191
	Palestinian	*THRB*	3.23E-06
		SLCO1C1	0.00086115
Europe	Basque	*SLCO1C1*	0.00035174
		TSHR	0.00085462
	BergamoItalian-Tuscan	*SLCO1C1*	7.78E-05
	French	*SULT6B1*	0.00072794
		SLC16A10	0.00072794
		IYD	0.00075574
	Orcadian	*SLC5A5*	9.83E-05
		THRB	0.00059697
		SULT6B1	0.00066424
	Russian	*SLC16A10*	0.00026039
		SECISBP2	0.00040399
		SULT6B1	0.00040441
	Sardinian	*THRB*	2.94E-05
		SECISBP2	3.27E-05
Central South Asia	Balochi	*SLCO1C1*	0.00065848
	Brahui	*SLC16A2*	1.51E-05
		SLCO1C1	0.00061849
	Burusho	*SLCO1C1*	0.00012304
		THRB	0.0004254
		SLC16A10	0.00094037
	Hazara	*SLCO1C1*	0.00076257
	Kalash	*THRB*	0.00019232
		SLC16A10	0.00079475

	Makrani	*DIO2*	0.00028549
	Pathan	*SLC16A10*	0.00069635
	Sindhi	*THRB*	0.00063874
		SLCO1C1	0.0009536
East Asia	Dai-Lahu	*THRB*	0.0008572
	Han	*TRIP4*	0.00013754
	Han	*SLCO1C1*	0.00013754
	Oroqen-Hezhen-Daur	*SLCO1C1*	0.00023563
	Naxi-Yi	*THRB*	0.00082179
	She-Miao-Tujia	*TTR*	0.0003107
	Xibo-Mongolian	*THRB*	0.00052523
		THRA	0.00052917
	Yakut	*DIO2*	0.00050707
		IYD	0.00072576
Americas	Maya	*TSHR*	0.00027492
		THRA	0.00027492
		THRB	0.00031588
		TRIP4	0.00036477
		SULT6B1	0.00081017
	Surui-Karitiana	*THRB*	0.00041299
Oceania	Papuan	*DIO2*	0.00027674

***Table S4.13**: The Iodine-associated genes within the 0.1% tail, as indicated by the F_{ST} selection values for each population.* *Ordered by most significant.*

Region	Population	Gene	F_{ST} $P-value$
Africa	Bantu-speaking	*THRB*	0.00025396
		SLC16A10	0.00081068
	Biaka	*THRB*	0.00079522
	Mandenka	*TSHR*	8.17E-05
		IYD	0.00043594
		SLCO1C1	0.000744
	Mbuti	*TRIP4*	3.96E-05
		THRB	0.00020352
		THRA	0.00070607
	San	*TSHR*	0.00080396
Middle-East	Bedouin	*TSHR*	0.00059462
	Druze	*TSHR*	0.00037124
	Mozabite	*THRB*	0.00063856
		TSHR	0.00066246
	Palestinian	*TSHR*	7.98E-05
		DIO1	0.0003277
		DIO2	0.00046738
Europe	Adygei	*TSHR*	0.00023855
		THRB	0.00041932
		SLCO1C1	0.00096203
	Basque	*DIO1*	0.00070659
		TSHR	0.00079007
	BergamoItalian-Tuscan	*DIO2*	0.00068438
	French	*DIO2*	0.00063582
		DIO1	0.0006635
		TSHR	0.0009121
	Orcadian	*THRB*	0.00044048
	Russian	*TSHR*	0.00073514
		TTR	0.00088984

		DIO1	0.00092937
Central-South Asia	Balochi	*THRB*	0.00064082
		TSHR	0.00080388
	Brahui	*TSHR*	0.00031802
		DIO2	0.00064879
	Burusho	*DIO1*	0.00013009
		DIO2	0.00039795
		THRA	0.0006907
	Hazara	*DIO1*	0.00041576
		TSHR	0.00045968
		SLCO1C1	0.00078412
	Kalash	*TSHR*	0.00016414
		DIO1	0.00032655
		DIO2	0.00058909
	Makrani	*DIO2*	0.00042861
		THRB	0.00054441
	Pathan	*DIO1*	0.00027736
		TSHR	0.00029277
		SLCO1C1	0.00097803
	Sindhi	*DIO1*	0.00046362
		SLCO1C1	0.00080735
	Uygur	*SLCO1C1*	0.00019682
		THRB	0.00021962
		TSHR	0.00044421
		DIO1	0.00057239
		DIO2	0.00062677
East Asia	Han	*DIO1*	0.00024443
		TSHR	0.00044923
		THRB	0.00066537
	Japanese	*DIO1*	0.00047758
		SULT6B1	0.00048852
		TSHR	0.00055913
		SLCO1C1	0.00075499
	Oroqen-Hezhen-Daur	*DIO1*	0.00041683
		TSHR	0.00049137
	NorthernHan-Tu	*DIO1*	0.00035868
		TSHR	0.00075354
	She-Miao-Tujia	*TSHR*	0.00056207
	Xibo-Mongolian	*SLCO1C1*	0.00093924
	Yakut	*DIO1*	0.0002481
		SULT6B1	0.00090321
Americas	Maya	*DIO1*	0.00062126
		TSHR	0.00062176
		DIO2	0.00067055
		TRIP4	0.00083053
	Pima	*SLCO1C1*	0.00018886
		THRB	0.00018886
		DIO2	0.00086237
	Surui-Karitiana	*THRB*	0.00012002
		TSHR	0.00023151
Oceania	Bougainville	*DIO2*	0.00034338

Figures

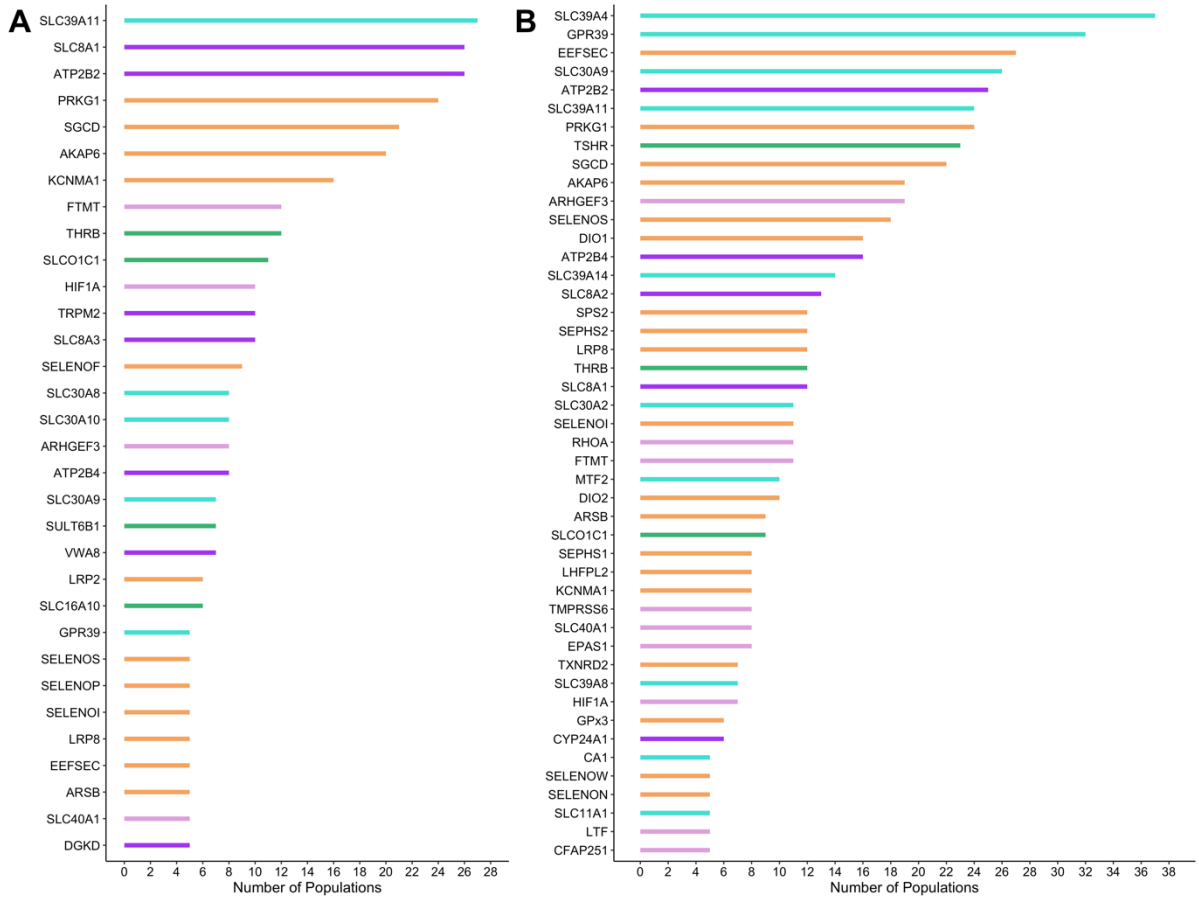


Figure S4.1: ZCSII-associated genes showing repeated signatures in the 0.1% tail. *Shown for* ***A)*** *Relate or* ***B)*** F_{ST} *selection values, with the number of populations showing such signatures given by the x-axis.*

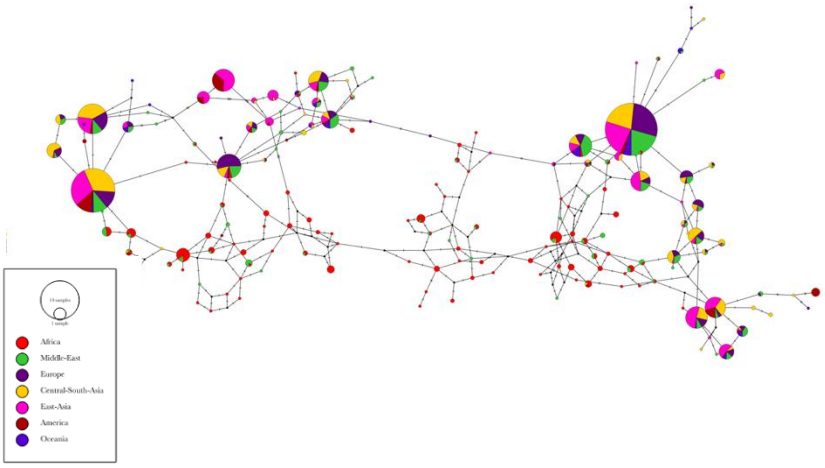


Figure S4.2: Haplotype network built from the 20kb region surrounding the chr8:144414297 SNP of SLC39A4

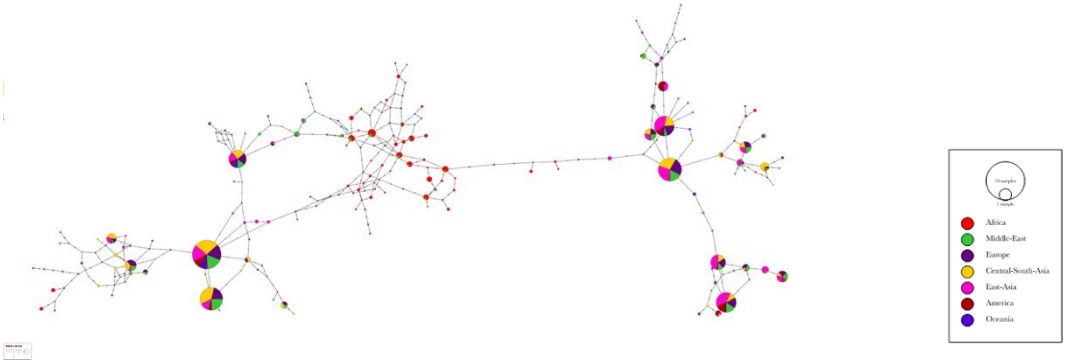


Figure S4.3: Haplotype network built from the 20kb region surrounding the chr2:132638916 SNP of GPR39

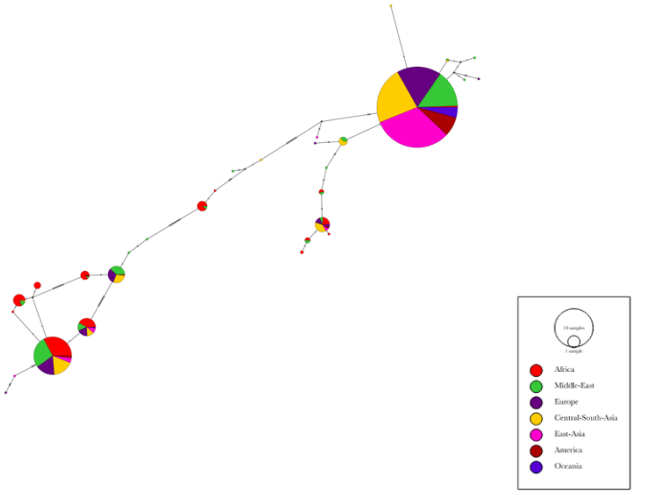


Figure S4.4: Haplotype network built from the 20kb region surrounding the chr4:42004040 SNP of SLC30A9

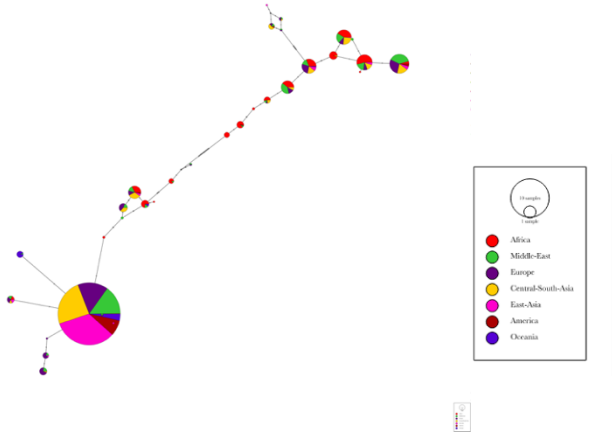


Figure S4.5: Haplotype network built from the 20kb region surrounding the chr4:42031397 SNP of SLC30A9

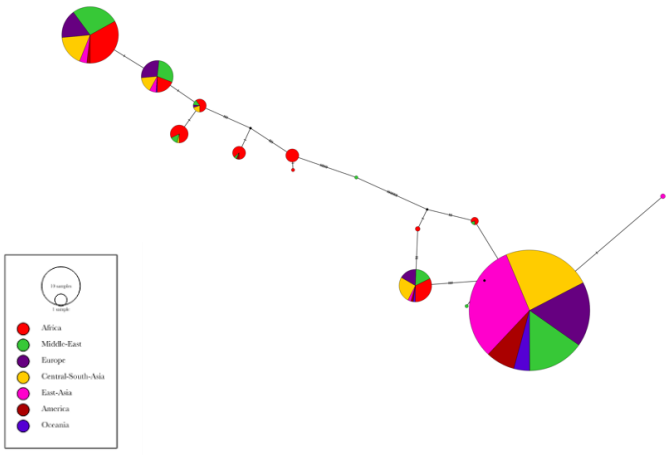


Figure S4.6: Haplotype network built from the 20kb region surrounding the chr4:42066213 SNP of SLC30A9

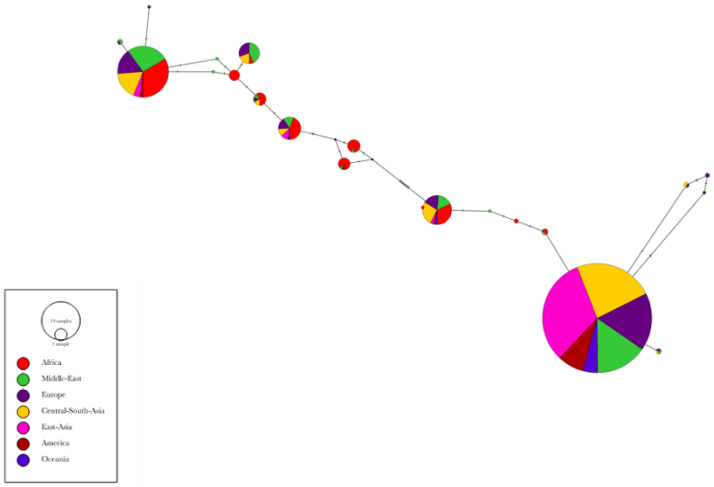


Figure S4.7: Haplotype network built from the 20kb region surrounding the chr4:42093983 SNP of SLC30A9

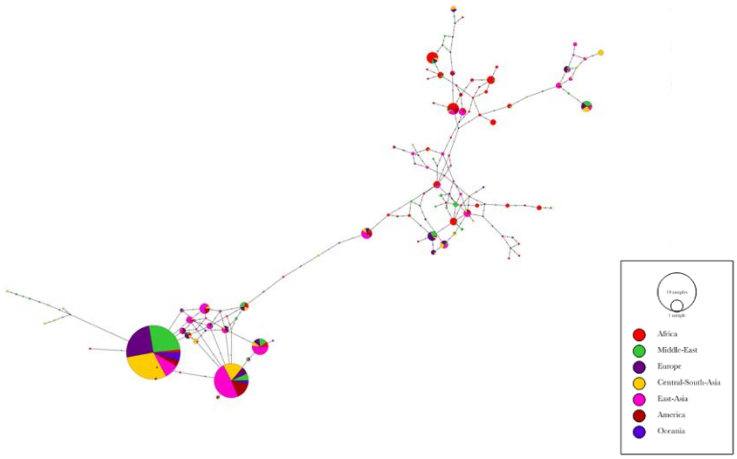


Figure S4.8: Haplotype network built from the 10kb region surrounding the chr17:73010373 SNP of SLC39A11

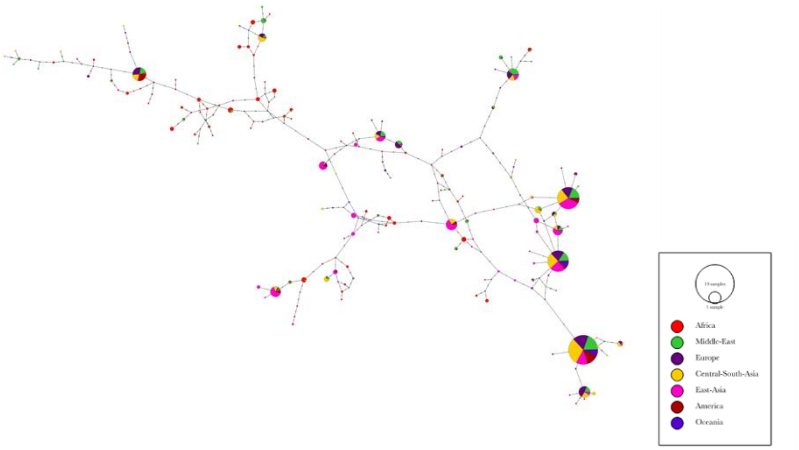


***Figure S4.9**: Haplotype network built from the 10kb region surrounding the chr17:72716374 SNP of SLC39A11*

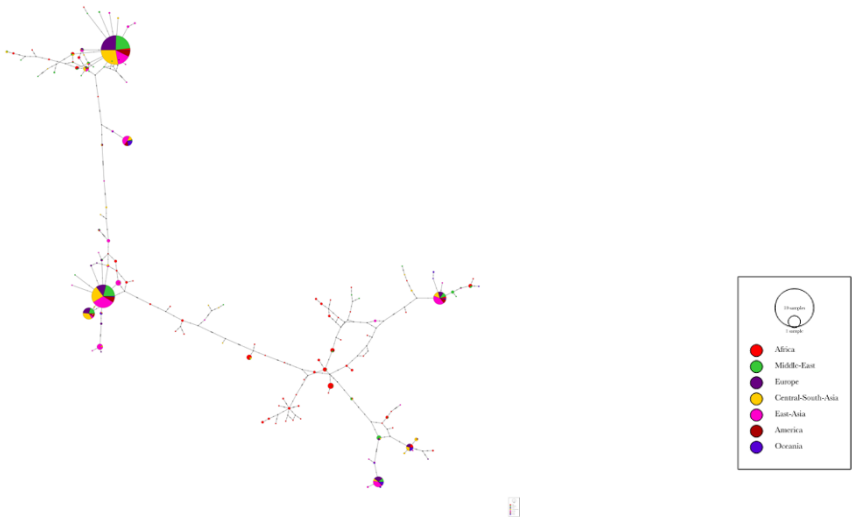


***Figure S4.10**: Haplotype network built from the 20kb region surrounding the chr8:22404076 SNP of SLC39A14*

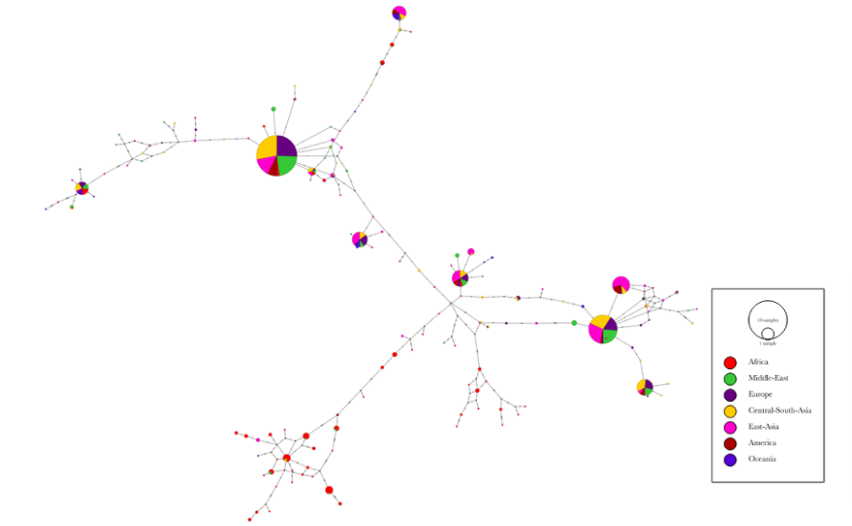


***Figure S4.11**: Haplotype network built from the 20kb region surrounding the chr8:22416174 SNP of SLC39A14*

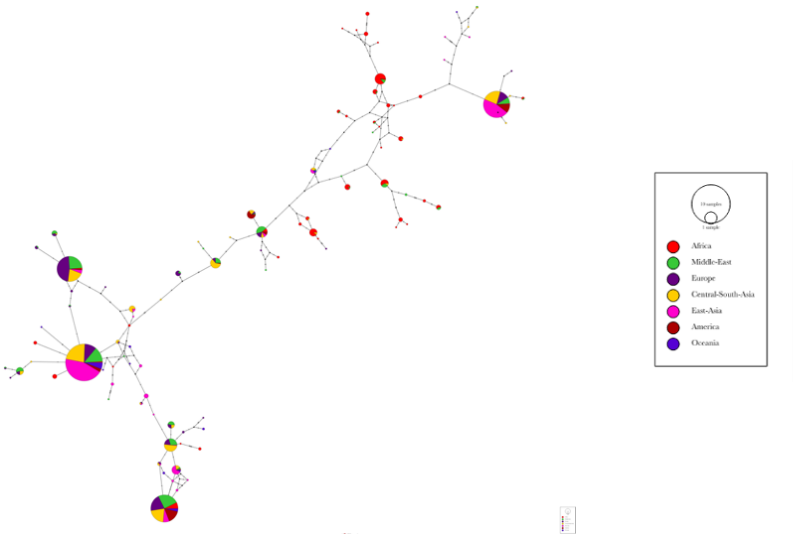


***Figure S4.12**: Haplotype network built from the 20kb region surrounding the chr3:10453703 SNP of ATP2B2*

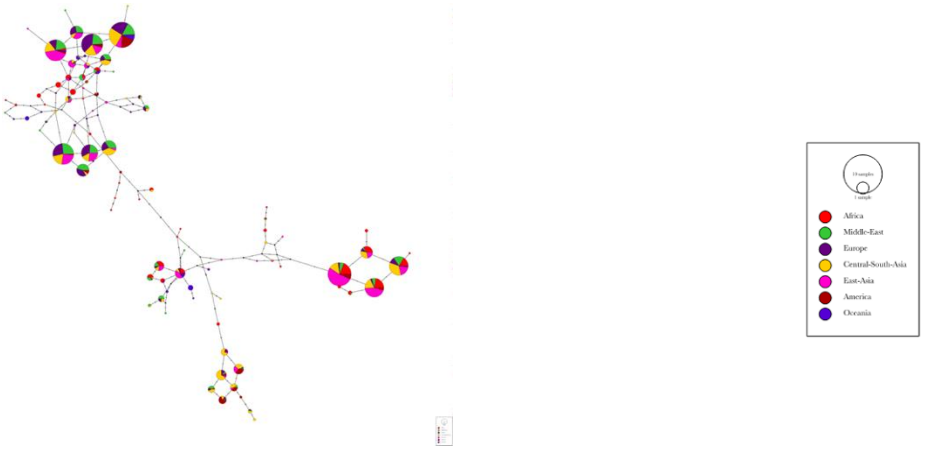


***Figure S4.13**: Haplotype network built from the 10kb region surrounding the chr3:10636328 SNP of ATP2B2*

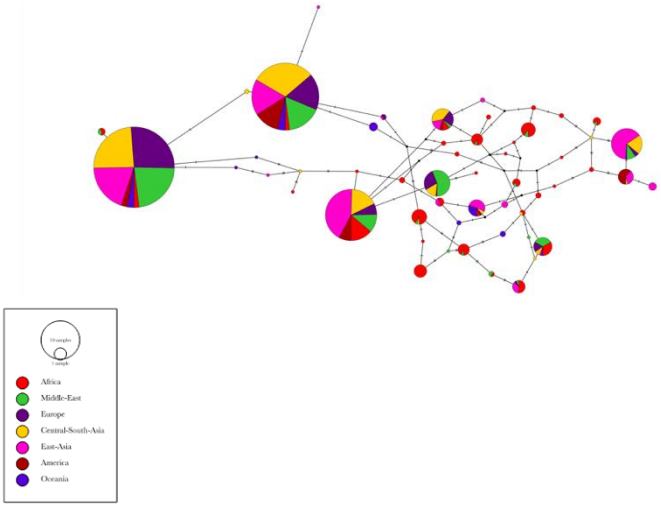


***Figure S4.14**: Haplotype network built from the 20kb region surrounding the chr1:203648263 SNP of ATP2B4*

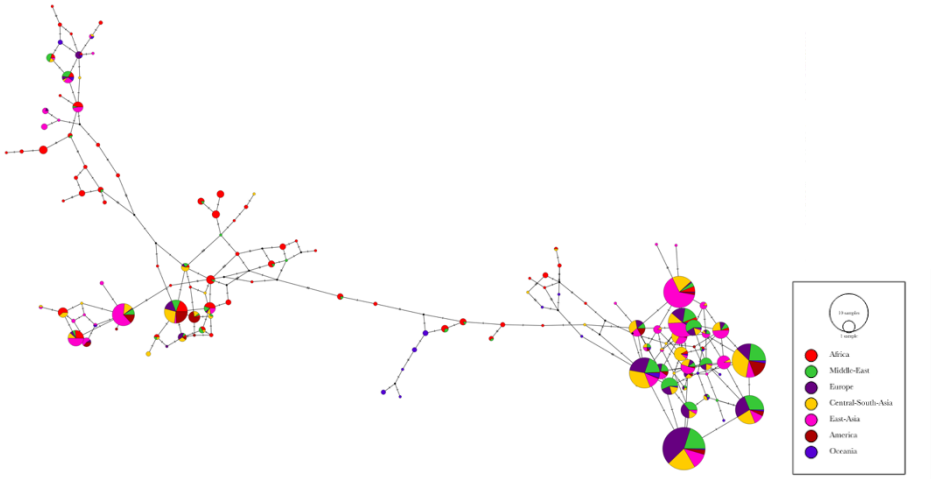


***Figure S4.15**: Haplotype network built from the 10kb region surrounding the chr1:203667951 SNP of ATP2B4*

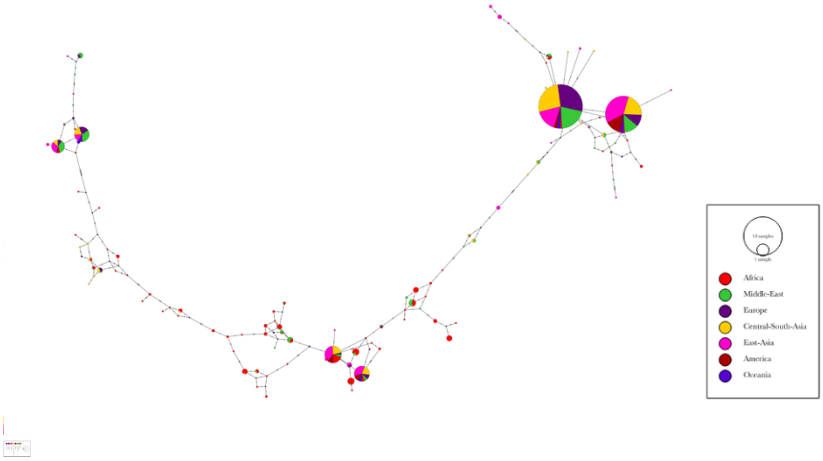


***Figure S4.16**: Haplotype network built from the 20kb region surrounding the chr19:47428756 SNP of SLC8A2*

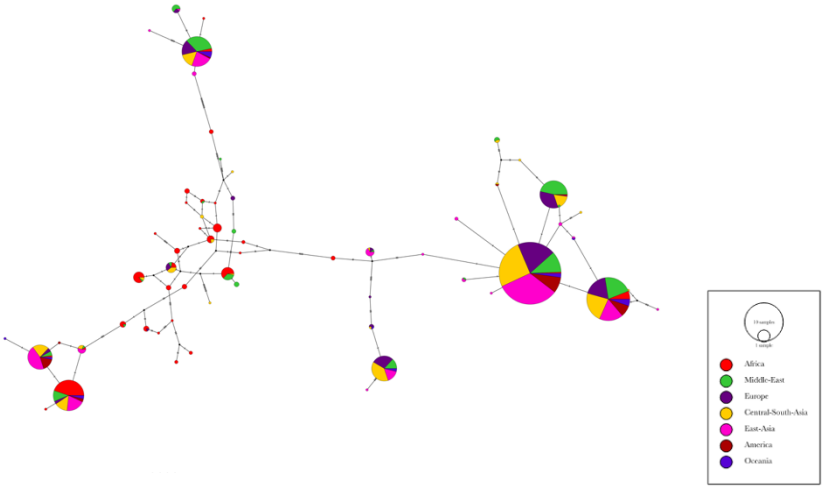


***Figure S4.17**: Haplotype network built from the 20kb region surrounding the chr19:47437107 SNP of SLC8A2*

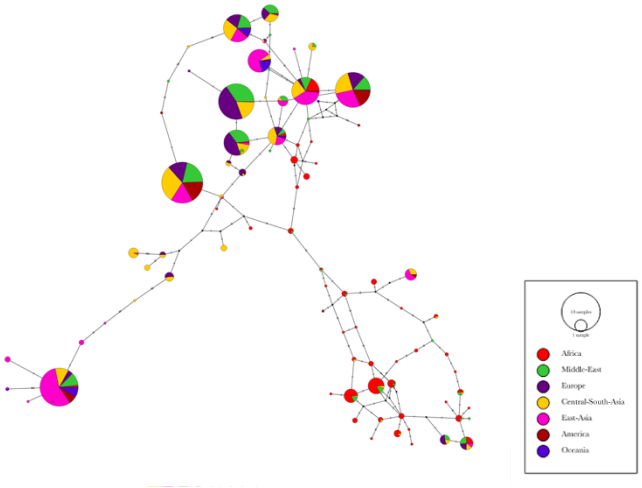


***Figure S4.18**: Haplotype network built from the 20kb region surrounding the chr3:128412869 SNP of EEFSEC*

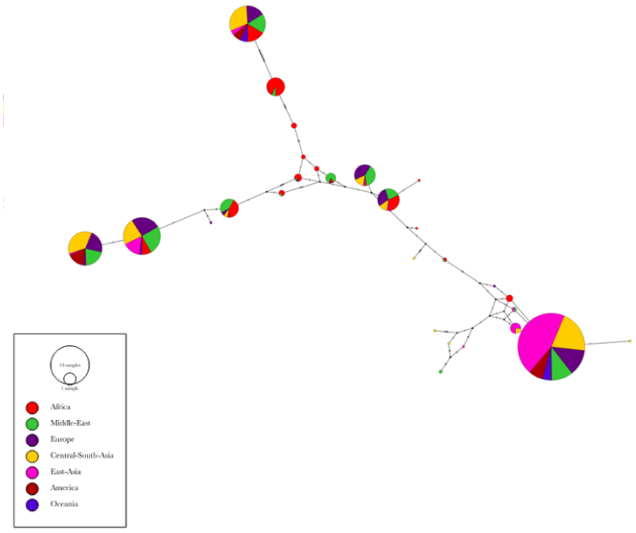


***Figure S4.19**: Haplotype network built from the 20kb region surrounding the chr10:51576270 SNP of PRKG1*

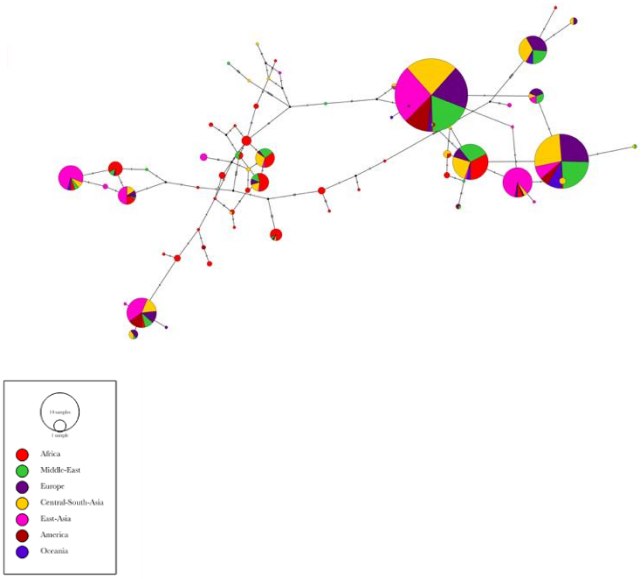


Figure S4.20: Haplotype network built from the 20kb region surrounding the chr10:51471686 SNP of PRKG1

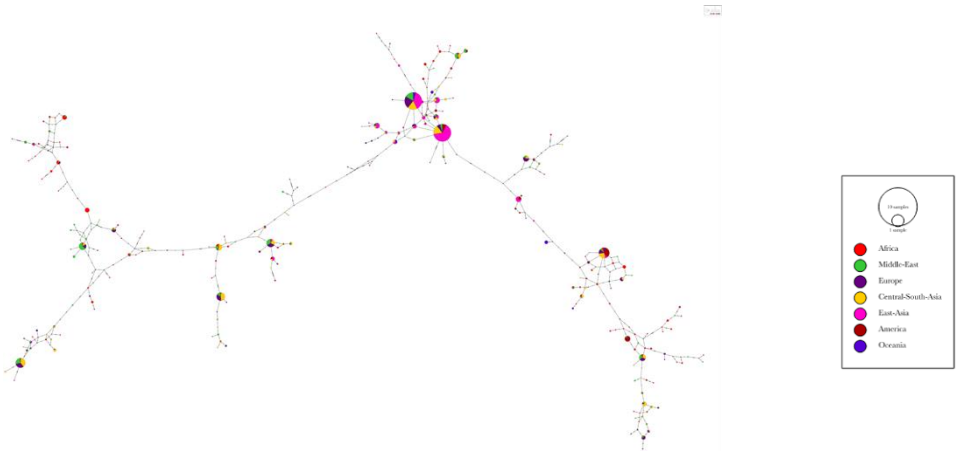


Figure S4.21: Haplotype network built from the 20kb region surrounding the chr5:156708844 SNP of SGCD

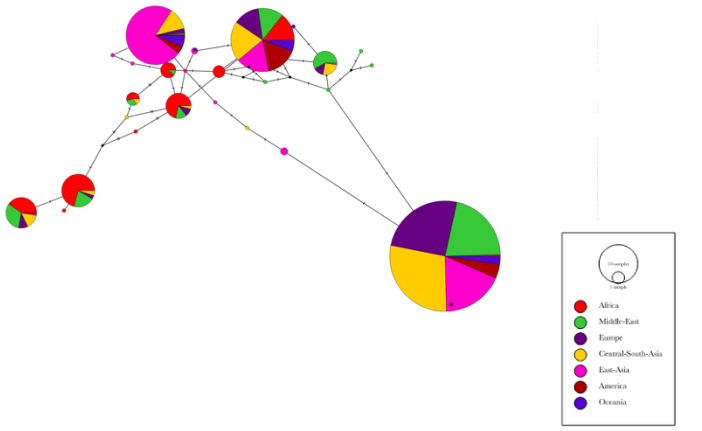


***Figure S4.22**: Haplotype network built from the 20kb region surrounding the chr5:156057959 SNP of SGCD*

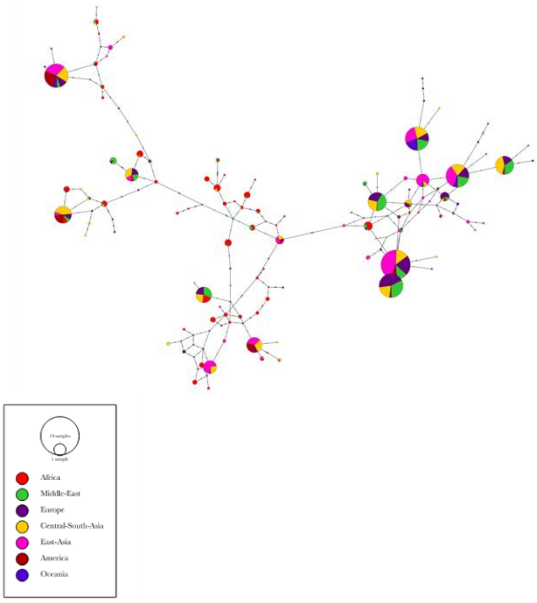


***Figure S4.23**: Haplotype network built from the 20kb region surrounding the chr14:32542441 SNP of AKAP6*

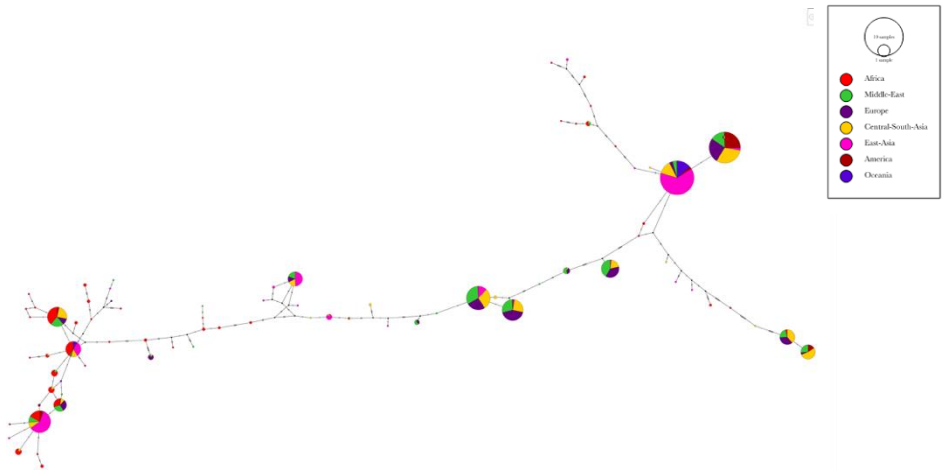


Figure S4.24: Haplotype network built from the 20kb region surrounding the chr14: 32446036 SNP of AKAP6

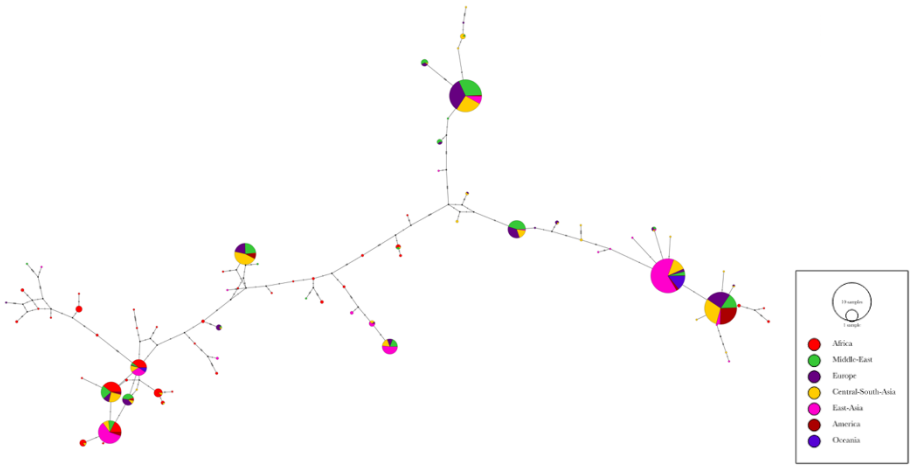


Figure S4.25: Haplotype network built from the 20kb region surrounding the chr14: 32453376 SNP of AKAP6

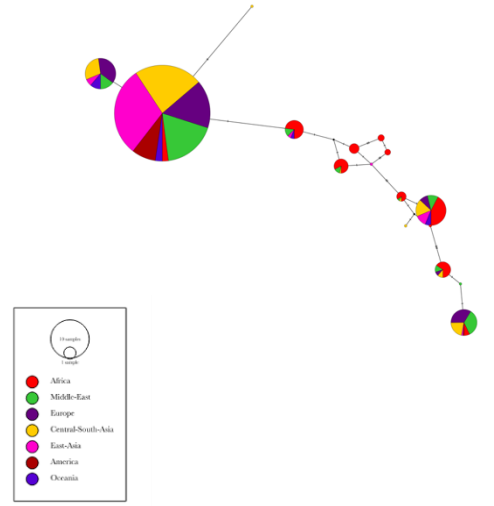


Figure S4.26: Haplotype network built from the 20kb region surrounding the chr1:53920598 SNP of DIO1

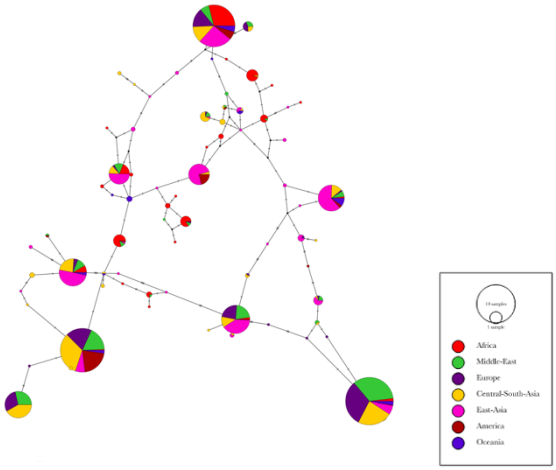


Figure S4.27: Haplotype network built from the 20kb region surrounding the chr3:56761998 SNP of ARHGEF3

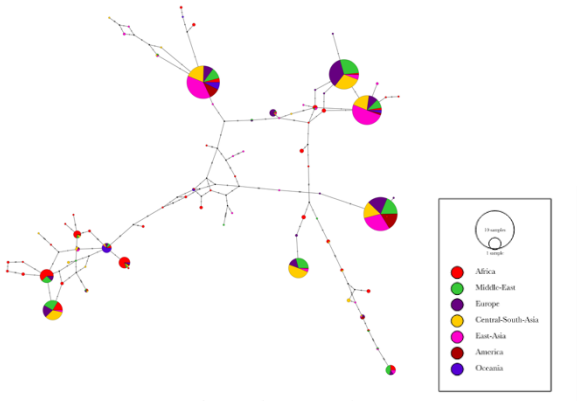


***Figure S4.28**: Haplotype network built from the 20kb region surrounding the chr14:80962759 SNP of TSHR*

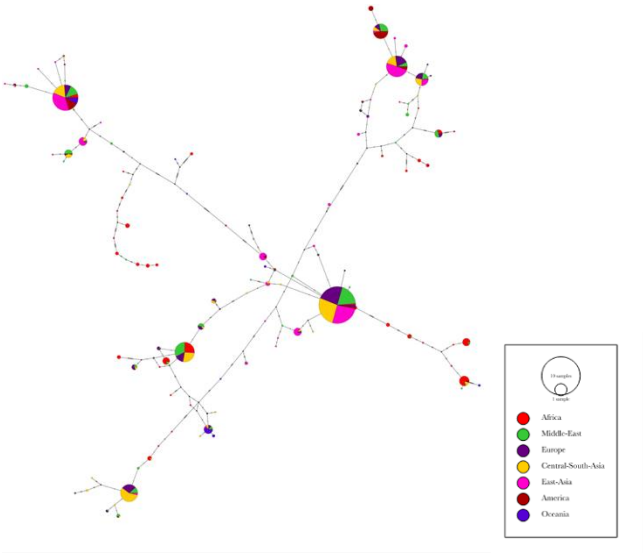


***Figure S4.29**: Haplotype network built from the 20kb region surrounding the chr14:81006112 SNP of TSHR*

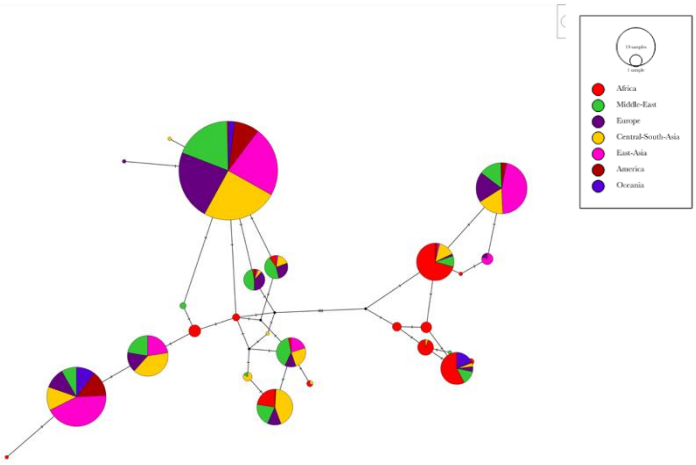


Figure S4.30: Haplotype network built from the 20kb region surrounding the chr14:81071140 SNP of TSHR

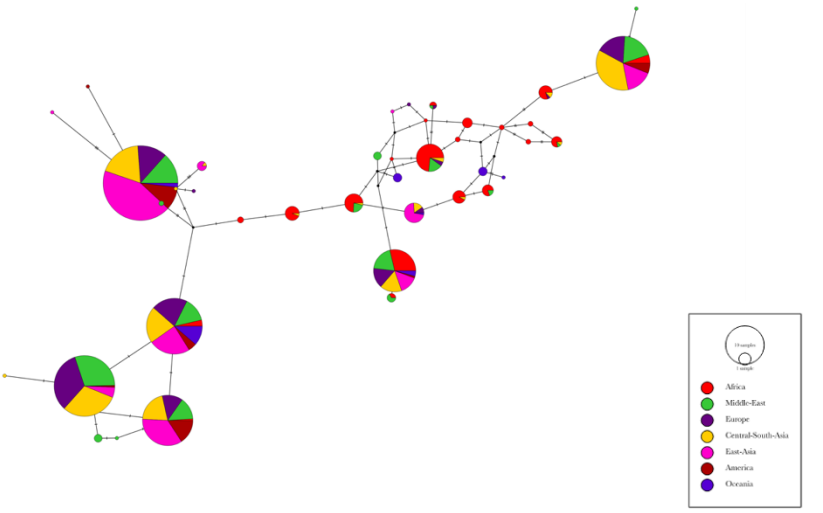


Figure S4.31: Haplotype network built from the 20kb region surrounding the chr3:24110895 SNP of THRB

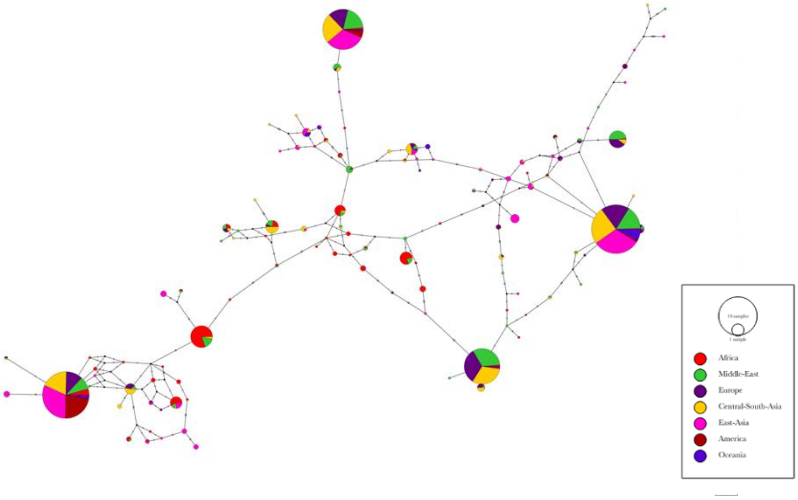


Figure S4.32: Haplotype network built from the 20kb region surrounding the chr3: 24342863 SNP of THRB

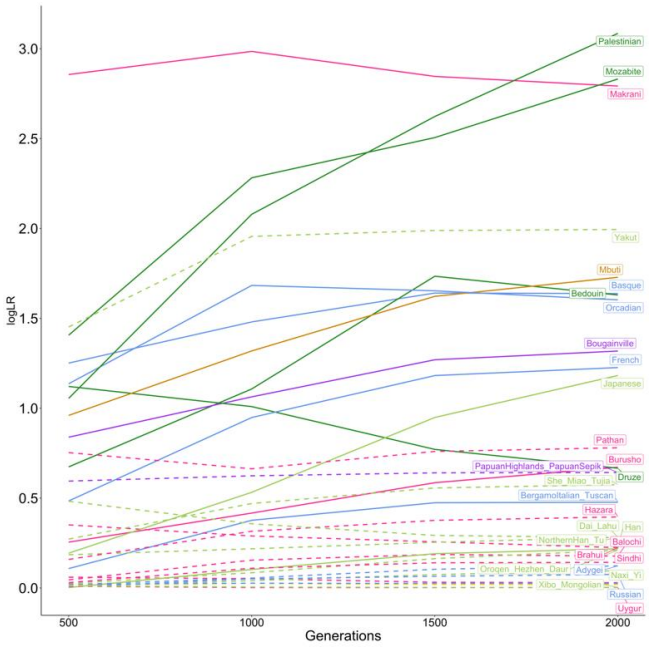


***Figure S4.33**: **Inferred log likelihood ratios for focal SNP of ATP2B2 (position: chr3:10456514)**. For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

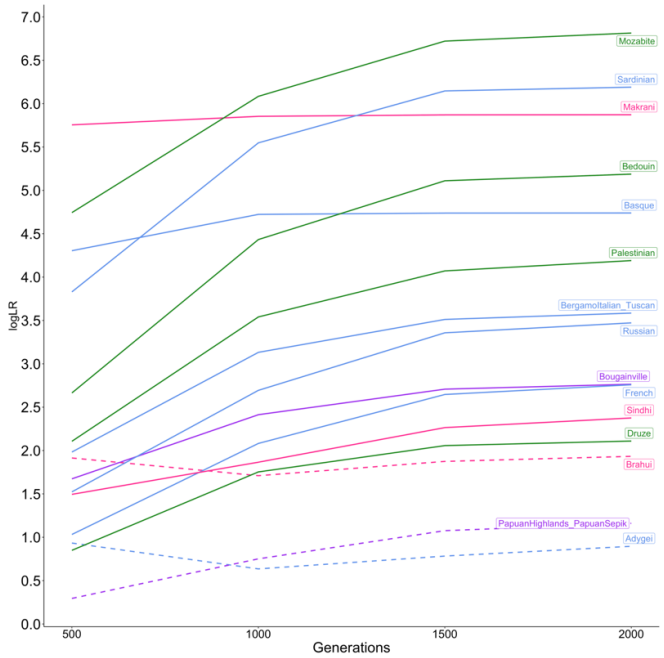


***Figure S4.34**: **Inferred log likelihoods ratios for focal SNP of ATP2B2 (position: chr3:10604833)**. For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

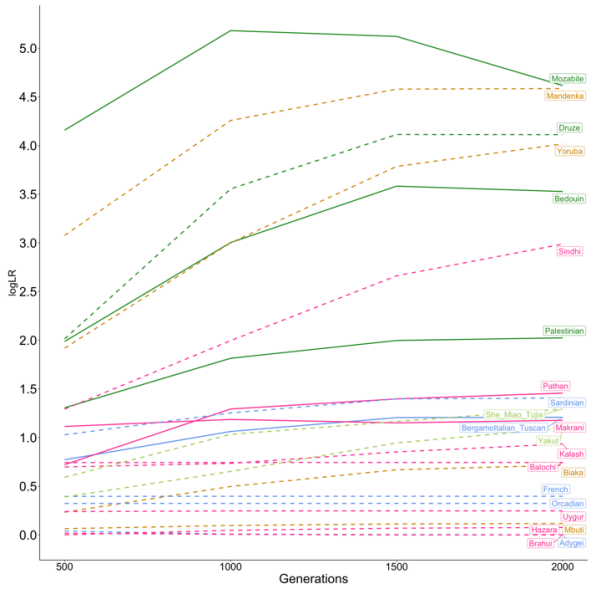


***Figure S4.35*: Inferred log likelihood ratios for focal SNP of ATP2B4 (position: chr1:203648263).** *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

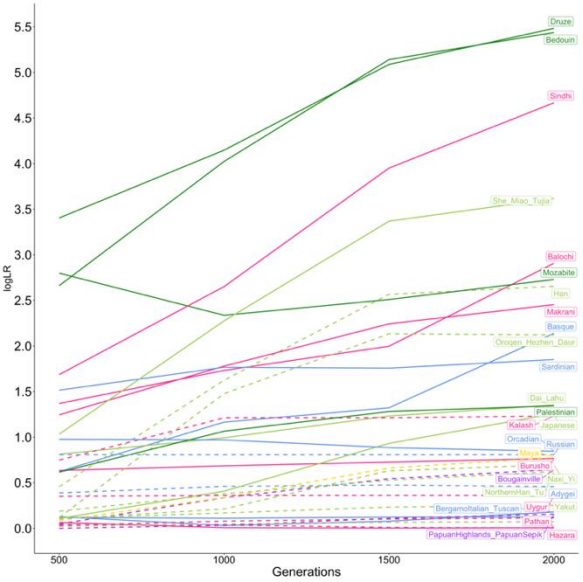


***Figure S4.36*: Inferred log likelihood ratios for focal SNP of ATP2P4 (position: chr1:203667951).** *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

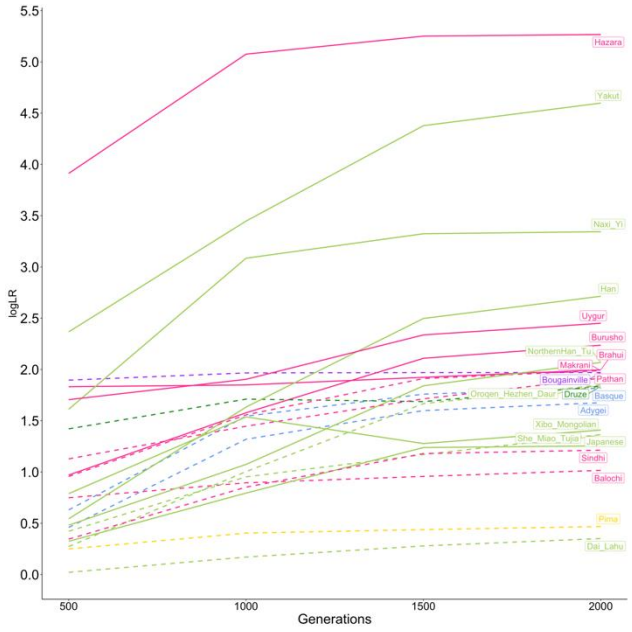


Figure S4.37: Inferred log likelihood ratios for focal SNP of SLC8A1 (position: chr2:40394610). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

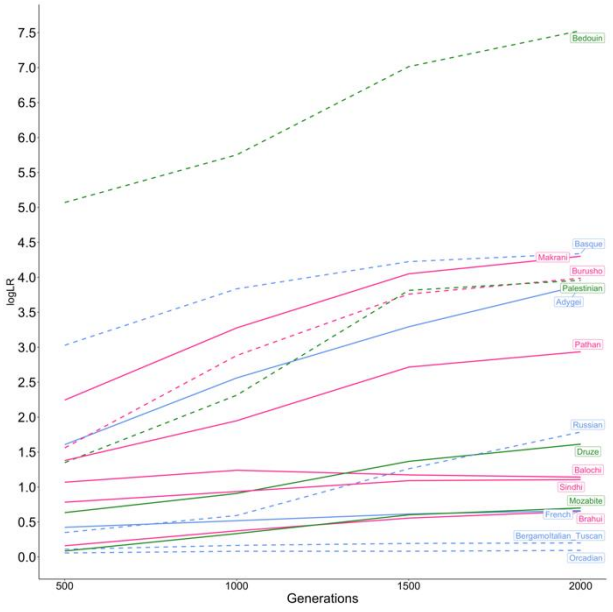


Figure S4.38: Inferred log likelihood ratios for focal SNP of SLC8A1 (position: chr2:40584510*). For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

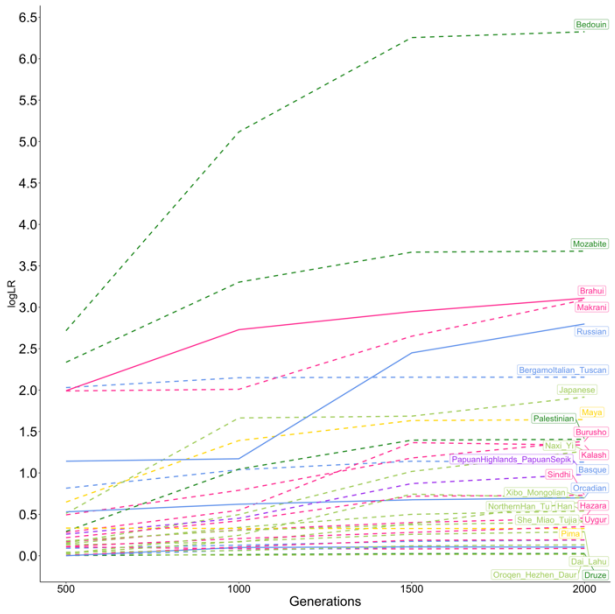


Figure S4.39*: *Inferred log likelihood ratios for focal SNP of SLC8A2 (position: chr19: 47428756). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

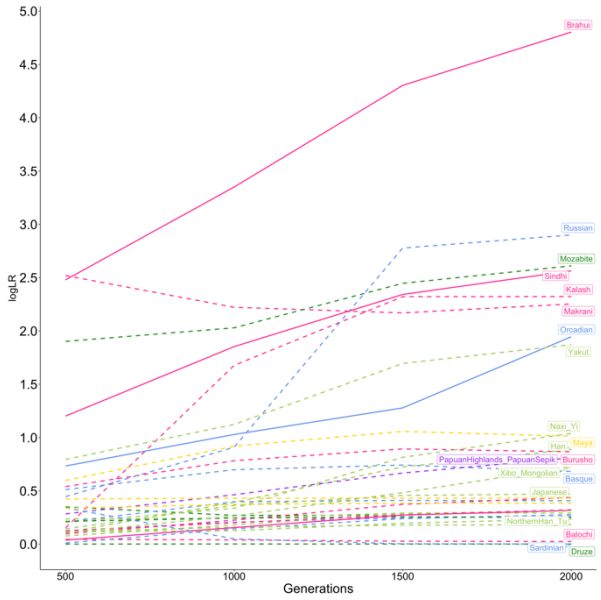


Figure S4.40*: *Inferred log likelihood ratios for focal SNP of SLC8A2 (position: chr19: 47437107). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

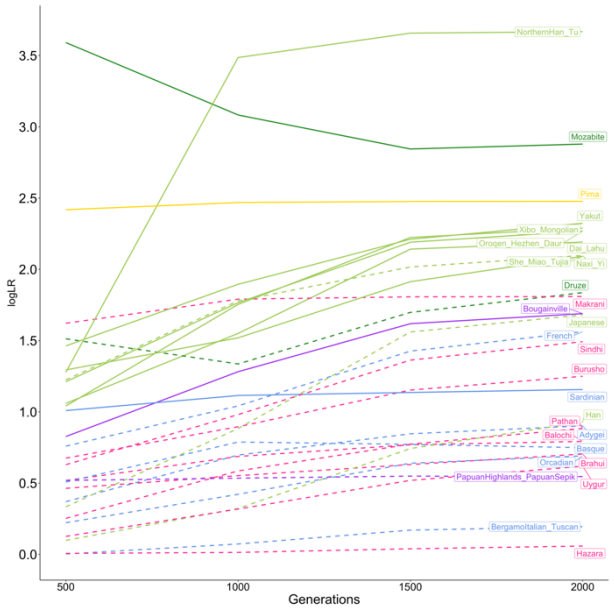


Figure S4.41: Inferred log likelihood ratios for focal SNP of SLC8A3 (position: chr14:70182346). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

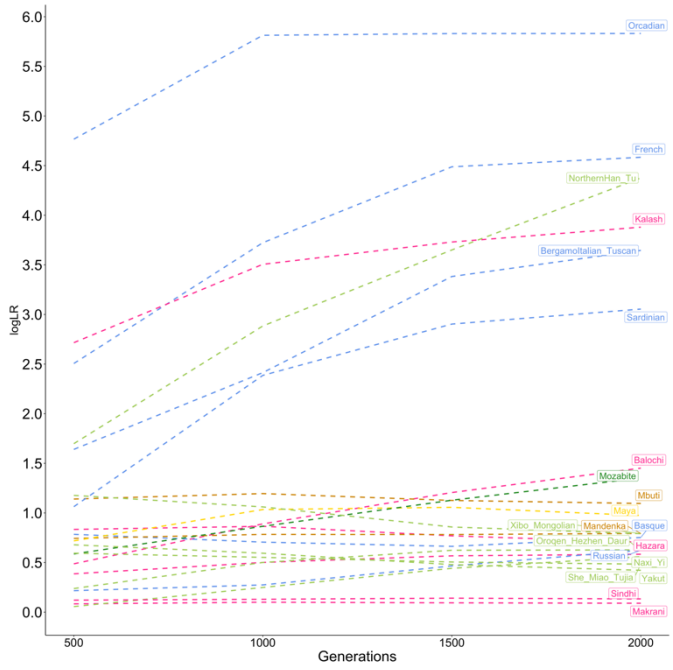


Figure S4.42: Inferred log likelihood ratios for focal SNP of SLC8A3 (position: chr14:70175561). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

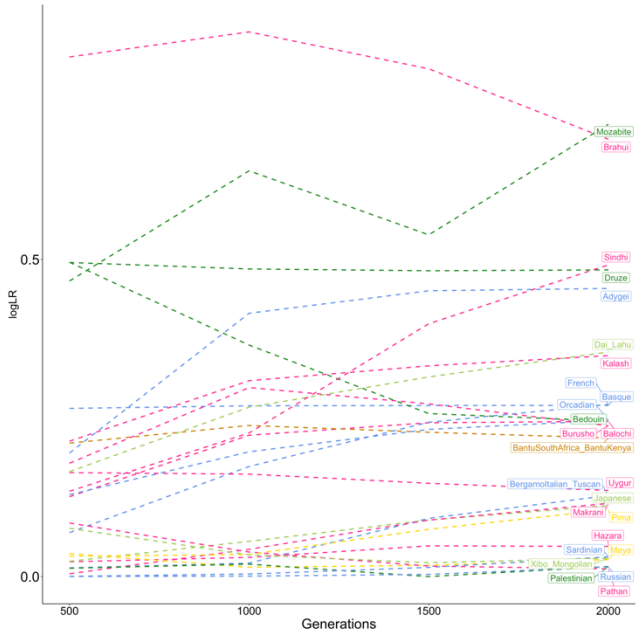


***Figure S4.43*: Inferred log likelihood ratios for focal SNP of ARHGEF3 (position: chr3:56761998).** *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

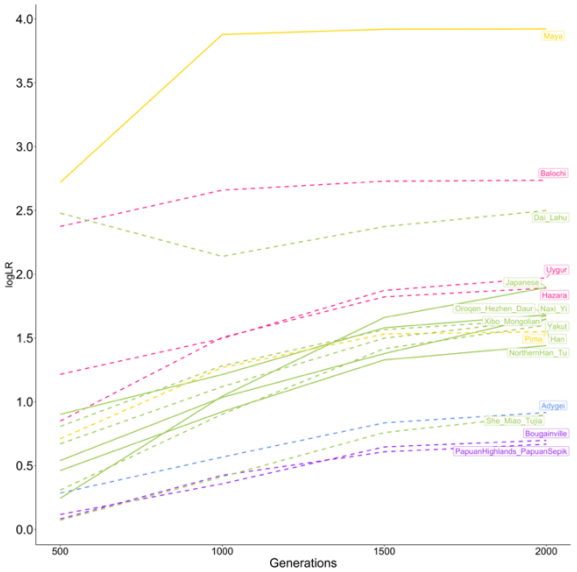


***Figure S4.44*: Inferred log likelihood ratios for focal SNP of ARHGEF3 (position: chr3:57043874).** *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

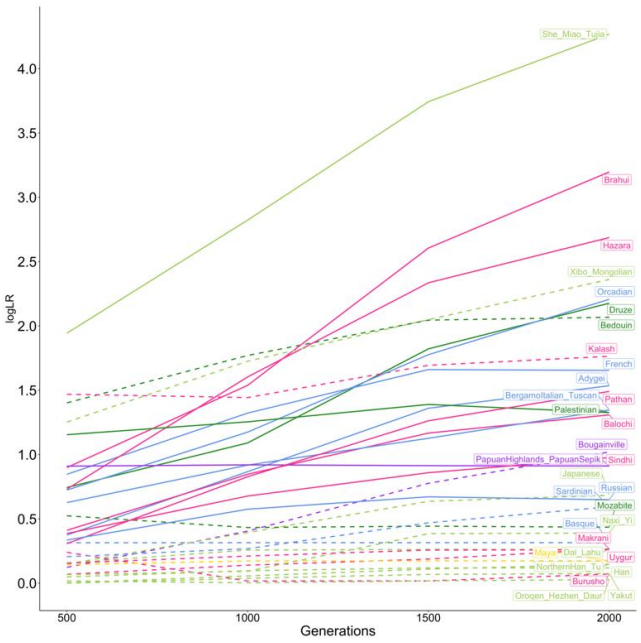


Figure S4.45*: *Inferred log likelihood ratios for focal SNP of HIF1A (position: chr14:61687412). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

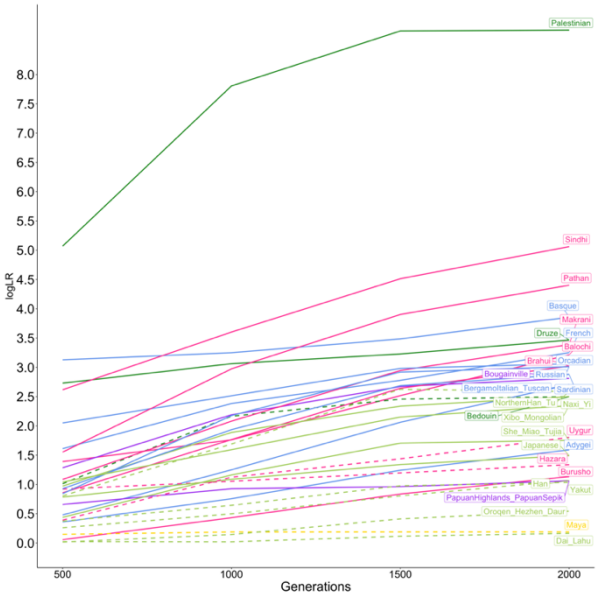


Figure S4.46*: *Inferred log likelihood ratios for focal SNP of HIF1A (position: chr14:61709502). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

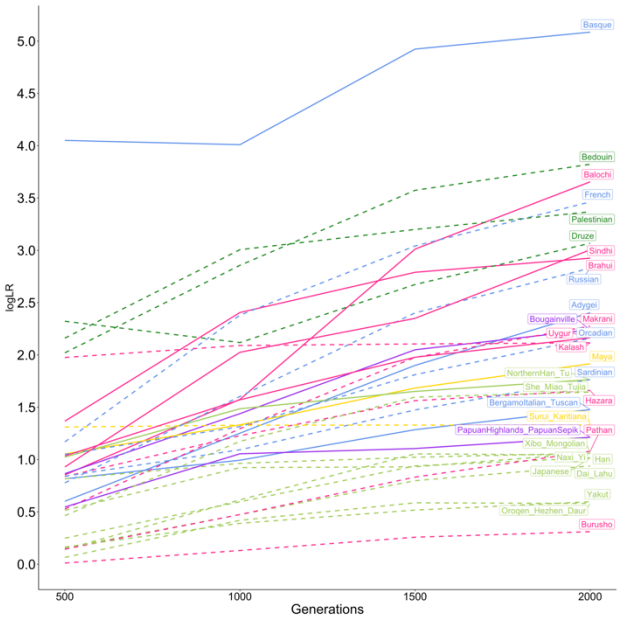


***Figure S4.47*: Inferred log likelihood ratios for focal SNP of HIF1A (position: chr14:61741756).** *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

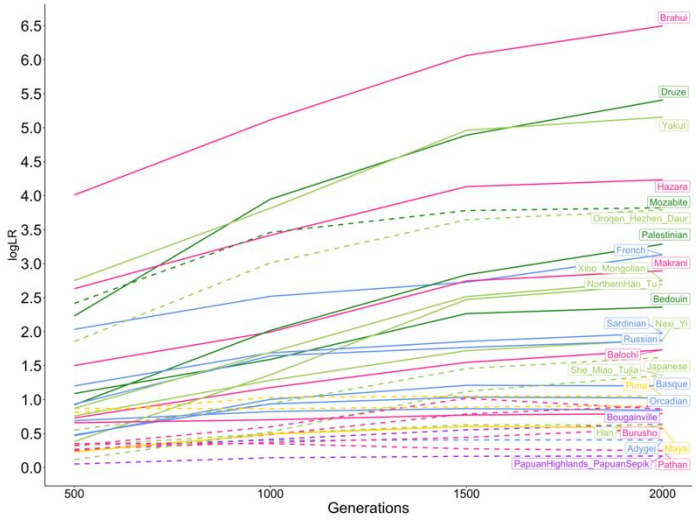


***Figure S4.48: Inferred log likelihood ratios for focal SNP of FTMT (position: chr5:121846819).** For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

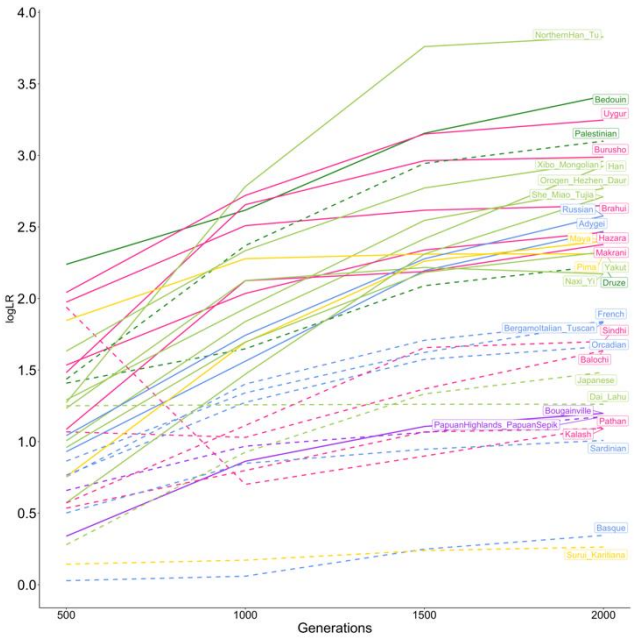


***Figure S4.49*: Inferred log likelihood ratios for focal SNP of FTMT (position: chr5:121853801).** *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

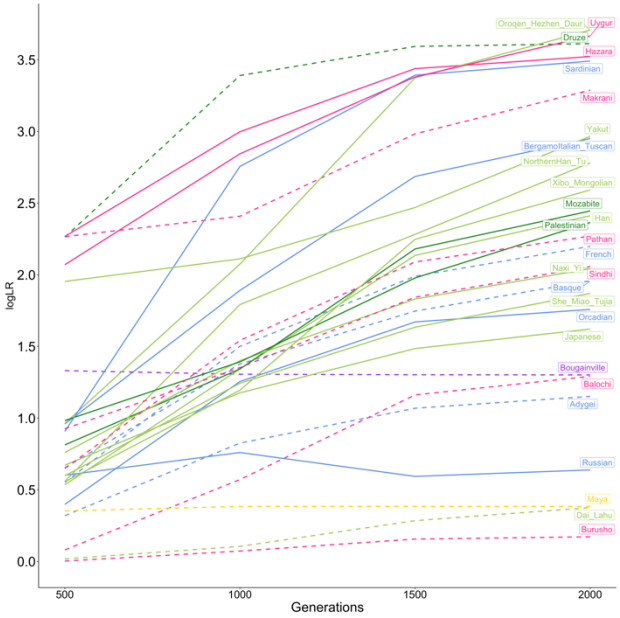


***Figure S4.50*: Inferred log likelihood ratios for focal SNP of SLC40A1 (position: chr2:189577426)**. *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

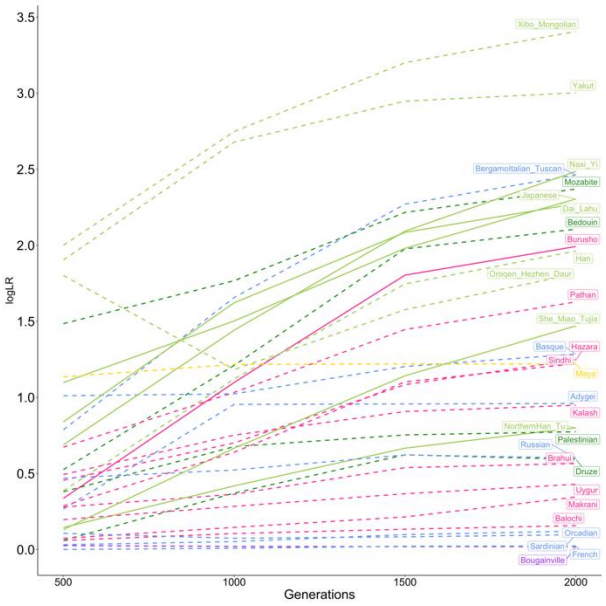


Figure S4.51: Inferred log likelihood ratios for focal SNP of SLC40A1 (position: chr2: 189591670). *For populations with selection values in the 5% tail for that SNP according to either Relate or F_{ST} (dashed lines) or according to both selection methods (solid lines). Colours represent metapopulations: blue = Europe; dark-green = Middle-East; pink = Central-South Asia; light-green = East Asia; yellow = America; purple = Oceania; dark-orange = Africa.*

Chapter 5: Supplementary Material

Tables

***Table S5.1: List of mammalian GPX coding sequences used for this study**. Latin names given in the GPX6$_{Sec/Cys}$ row.*

GPX Protein	Coding Sequences Included
GPX6$_{Sec/Cys}$	Bushbaby (*Otolemur garnettii*), Cat (*Felis catus*), Chimpanzee (*Pan troglodytes*), Chinese hamster (*Cricetulus griseus*), Cow (*Bos taurus*), Elephant (*Loxodonta africana*), Golden hamster (*Mesocricetus auratus*), Guinea pig (*Cavia porcellus*), Horse (*Equus caballus*), Human (*Homo sapiens)*, Jerboa (*Jaculus jaculus*), Kangaroo rat (*Dipodomys ordii*), Macaque (*Macaca mulatta)*, Marmoset (*Callithrix jacchus*), Mouse (*Mus musculus*), Pig (*Sus scrofa*), Rabbit (*Oyctolagus cuniculus*), Rat (*Rattus norvegicus*), Squirrel (*Ictidomys tridecemlineatus*), Squirrel monkey (*Saimiri boliviensis*), Tarsier (*Carlito syrichta*), Walrus (*Odobenus rosmarus*).
GPX1$_{Sec}$	Bushbaby, Cat, Chimpanzee, Chinese hamster, Cow, Elephant, Golden hamster, Human, Jerboa, Kangaroo rat, Macaque, Mouse, Pig, Rabbit, Rat, Squirrel, Squirrel monkey, Tarsier
GPX2$_{Sec}$	Bushbaby, Cat, Chimpanzee, Chinese hamster, Cow, Dog, Elephant, Golden hamster, Guinea pig, Horse, Human, Jerboa, Kangaroo rat, Macaque, Marmoset, Mouse, Pig, Rabbit, Rat, Squirrel, Squirrel monkey
GPX3$_{Sec}$	Bushbaby, Cat, Chimpanzee, Chinese hamster, Cow, Dog, Elephant, Golden hamster, Guinea pig, Horse, Human, Jerboa, Macaque, Marmoset, Mouse, Pig, Rabbit, Rat, Squirrel, Tarsier
GPX4$_{Sec}$	Cat, Chimpanzee, Chinese hamster, Cow, Elephant, Gibbon, Golden hamster, Guinea pig, Horse, Human, Kangaroo rat, Macaque, Mouse, Mouse lemur, Pig, Rat, Squirrel, Squirrel monkey
GPX5$_{Cys}$	Bushbaby, Cat, Chimpanzee, Chinese hamster, Cow, Dog, Elephant, Golden hamster, Guinea pig, Horse, Human, Jerboa, Kangaroo rat, Macaque, Marmoset, Mouse, Pig, Rabbit, Rat, Squirrel, Squirrel monkey, Tarsier, Walrus
GPX7$_{Cys}$	Bushbaby, Chimpanzee, Chinese hamster, Cow, Dog, Elephant, Golden hamster, Guinea pig, Horse, Human, Jerboa, Kangaroo rat, Macaque, Marmoset, Mouse, Rabbit, Rat, Squirrel, Tarsier
GPX8$_{Cys}$	Bushbaby, Cat, Chimpanzee, Chinese hamster, Cow, Elephant, Golden hamster, Guinea pig, Horse, Human, Jerboa, Kangaroo rat, Macaque, Marmoset, Mouse, Panda, Pig, Rabbit, Rat, Squirrel, Squirrel monkey, Tarsier

Table S5.2: dN/dS ratios for the GPX family and protein regions*. Given for lineages where GPX6 has Sec (Fig. 1, solid red branches), exchanged Sec for Cys (Fig. 1, dashed green branches) or inherited Cys (Fig. 1, solid green branches). In some lineages for which PAML estimates very few synonymous changes compared to the non-synonymous changes, unnaturally large dN/dS values can occur; these are marked with a #. The dN/dS ratio for all branches is the null hypothesis (one ratio for all branches) used in the likelihood ratio test contrasting the two previous ones. P-values are obtained based on a χ^2 distribution with d.f=2. In bold, significant P-values.*

Protein	Region	dN/dS in branches where GPX6 has				P-value
		Sec	Cys after Sec was lost	Inherited Cys	All	
GPX1$_{Sec}$	Full length	0.080	0.045	0.087	0.074	0.115
	N-terminus	**0.043**	**0.009**	**0.190**	**0.034**	**0.046**
	GPX	0.064	0.040	0.069	0.060	0.534
	C-terminus	0.085	0.052	0.114	0.081	0.328
GPX2$_{Sec}$	Full length	**0.069**	**0.029**	**0.041**	**0.055**	**0.024**
	N-terminus	0.032	0.001	0.001	0.032	0.999
	GPX	0.075	0.042	0.038	0.060	0.191
	C-terminus	0.055	0.017	0.048	0.043	0.100
GPX3$_{Sec}$	Full length	0.132	0.131	0.077	0.125	0.222
	N-terminus	**0.241**	**0.038**	**0.461**	**0.181**	**0.022**
	GPX	0.094	0.108	0.056	0.091	0.439
	C-terminus	0.105	0.195	0.058	0.114	0.161
GPX4$_{Sec}$	Full length	0.071	0.073	0.112	0.076	0.380
	N-terminus	0.108	0.018	0.123	0.082	0.264
	GPX	**0.062**	**0.007**	**0.203**	**0.061**	**$1x10^{-4}$**
	C-terminus	0.043	0.003	0.033	0.030	0.126
GPX5$_{Cys}$	Full length	0.294	0.258	0.429	0.305	0.061
	N-terminus	0.678	0.634	0.959	0.716	0.716
	GPX	0.233	0.145	0.219	0.212	0.227
	C-terminus	0.237	0.225	0.358	0.250	0.379
GPX7$_{Cys}$	Full length	0.141	0.086	0.157	0.137	0.377
	N-terminus	0.190	#999	0.005	0.122	0.070
	GPX	0.083	0.080	0.117	0.088	0.712
	C-terminus	0.185	0.080	0.224	0.178	0.242
GPX8$_{Cys}$	Full length	0.228	0.156	0.156	0.203	0.199
	N-terminus	0.169	0.195	0.112	0.158	0.775
	GPX	0.223	0.155	0.198	0.207	0.616
	C-terminus	0.217	0.161	0.104	0.194	0.486

Table S5.3: Foreground branches tested using the branch-site model. *The amino acid sites inferred as under selection using the Bayes Empirical Bayes inference listed when supported by a P-value < 0.05. P-values are obstained based on a χ^2 distribution with d.f=1. Posterior probabilities of selection are shown in parehtneses, in bold face when P > 0.9. All foreground branches used here are lineages where Sec was exchanged for Cys (Fig 1, dashed green branches)*

Region	Foreground branches	P-value	Sites under selection
GPX domain	Eumuroida, Rabbit, Primate, Cat, Walrus	**0.046**	45 (0.783); 46 (0.571); **50 (0.946); 52 (0.936); 56 (0.950);** 57 (0.513); **62 (0.995);** 63 (0.535); 70 (0.552); **74 (0.992);** 75 (0.573); **77 (0.975); 83 (0.995);** 91 (0.728); **110 (0.937); 126 (0.947)**; 143 (0.799); 149 (0.606)
Full protein	Eumuroida, Rabbit	**0.006**	16 (0.741); **45 (0.913);** 56 (0.895); **74 (0.992)**; **77 (0.942)**; **83 (0.992)**; 126 (0.822); 171 (0.784); 214 (0.573); 215 (0.668)
	Eumuroida, Rabbit, Primate	**0.008**	**16 (0.988)**; 45 (0.858); **56 (0.935)**; 63 (0.501); 73 (0.503); **74 (0.991)**; **77 (0.971)**; **83 (0.988)**; **110 (0.925)**; **126 (0.937)**; 149 (0.554); 171 (0.757); 189 (0.520); 207 (0.502); 212 (0.758); 214 (0.698); 215 (0.682); 216 (0.822)
	Eumuroida, Rabbit, Primate, Cat,	0.054	
	Eumuroida, Rabbit, Primate, Walurs	0.050	
	Eumuroida, Rabbit, Primate, Cat, Walrus	0.127	

Table S5.4: Convergent (sites change to the same amino acid) and pseudo-convergent (sites change to different amino acids) sites identified between GPX6$_{Cys}$ lineages by CONVERG2. *The two left-most columns are the branches between which convergent or pseudo-convergent sites are identified. The number gives the amino acid site where convergence is identified; the brackets represent the (ancestral amino acids on both branches, derived amino acids on both branches) in the order of the branches. Here, (SM-M) stands for the Squirrel monkey – marmoset internal branch, (GH-CH) stands for the Golden hamster – Chinese hamster internal branch and (rat-mouse) represents the rat-mouse internal branch. The branch names given in green indicate the branches upon which we have inferred the Sec-to-Cys exchange to have occurred. Convergent sites are in bold, all other sites are pseudo-convergent.*

(SM-M)	Eumuroida	15 (GG,SA)	166 (SK,NE)								
(SM-M)	Rabbit	15 (GG,SS)	34 (GG, NE)	62 (HH,PY)	91 (FF, LL)	110 (TK,AR)	166 (SK,NR)	212 (EE,AK)			
(SM-M)	Cat	26 (NK,DT)	**48 (NN,SS)**	62 (HH,PV)	90 (NP,SR)	91 (FF,LY)	166 (SS,ND)	190 (DD,HN)			
(SM-M)	Walrus	91 (FF,LS)									
(SM-M)	Golden hamster	48 (NN,ST)									
(SM-M)	Chinese hamster	62 (HH,PY)									
(SM-M)	(Rat-mouse)	48 (NN,SD)	90 (QN,PS)								
Eumuroida	Rabbit	15 (GG,AS)	45 (LL,KR)	**56 (KK,QQ)**	83 (AA,TR)	92 (GG,NS)	**143 (KK,NN)**	165 (SS,TP)	166(KK,E R)	**167 (QQ,HH)**	**171 (EE,D D)**
Eumuroida	Cat	**70 (TT,SS)**	**143 (KK,NN)**	165 (SS,TP)	166 (KS,ED)	207(KK,Q R)					
Eumuroida	Walrus	165 (SS,TA)									
Eumuroida	Squirrel monkey	48 (LL,KF)	**63 (VV,II)**								
Eumuroida	Marmoset	**70 (TT,SS)**									
Eumuroida	Rat	127 (FY,YF)									
Eumuroida	Chinese hamster	27 (MA,AE)									
Eumuroida	(Rat-mouse)	27 (MA,AS)	45 (LK,KN)	165 (ST,PT)							
Rabbit	Cat	62 (HH,YV)	91 (FF,LY)	**143 (KK,NN)**	**165 (SS,PP)**	166 (KS,RD)					
Rabbit	Walrus	51 (YY,DH)	91 (FF,LS)	165 (SP,PA)							
Rabbit	Squirrel monkey	45 (LL,FR)									
Rabbit	Marmoset	73 (GG,AS)	**208 (SS,AA)**								
Rabbit	Mouse	51 (YY,DF)	200 (KQ,AH)								
Rabbit	Rat	36 (TT,SA)	192 (VV,IA)								
Rabbit	Golden hamster	36 (TT,SA)									
Rabbit	Chinese hamster	**62 (HH,YY)**	189 (PP,TS)								
Rabbit	(Rat-mouse)	11 (PP,LS)	45 (LK,RN)	160 (ST,PP)	200 (KK,AQ)						
Rabbit	(GH-CH)	167 (QH,HY)	171 (ED,DN)								
Cat	Walrus	91 (FF,YS)	165 (SS,PA)								
Cat	Squirrel monkey	34 (KK,TN)	50 (EE,GD)								
Cat	Marmoset	**70 (TT,SS)**									
Cat	Golden hamster	48 (NN,TS)									
Cat	Chinese hamster	62 (HH,YV)									
Cat	(Rat-mouse)	48 (NN,SD)	90 (PQ,RP)	165 (ST,PP)							

Walrus	Mouse	51 (YY,HF)	54 (QQ,PN)
Walrus	(Rat-mouse)	165 (ST,AP)	
Walrus	(GH-CH)	**54 (QQ,PP)**	
Squirrel monkey	(Rat-mouse)	45 (LK,FN)	
Marmoset	Rat	94 (IT,TS)	
Marmoset	(Rat-mouse)	**94 (II,TT)**	
Mouse	(Rat-mouse)	200 (QK,HQ)	
Mouse	(GH-CH)	54 (QQ,NP)	
Rat	Golden hamster	**36 (TT,AA)**	
Rat	(Rat-mouse)	94 (TI,ST)	205 (IT,VI)
Golden hamster	(Rat-mouse)	48 (NN,TD)	
Chinese hamster	(Rat-mouse)	27 (AA,ES)	

Table S5.5: Convergent and pseudoconvergent sites in GPX1 between the GPX6$_{Cys}$ lineages, where the sequences were also available, as identified by CONVERG2*. The two left-most columns are the branches between which convergent or pseudo-convergent lineages are identified. The number gives the amino acid site where convergence is identified; the brackets represent (ancestral amino acids on both branches, derived amino acids on both branches) in the order of the branches. Here, (SM-M) stands for the Squirrel monkey – marmoset internal branch, (GH-CH) stands for the Golden hamster – Chinese hamster internal branch and (rat-mouse) represents the rat-mouse internal branch. The branch names given in green indicate the branches upon which we have inferred the Sec-to-Cys exchange in GPX6$_{Cys}$ to have occurred. Strict convergent sites are given in bold; all other sites are pseudo-convergent.*

Rabbit	Mouse	**177 (PP,SS)**	
Rabbit	Rat	**138 (AA,SS)**	**175 (QQ,KK)**
Rabbit	Golden hamster	**10 (SS,NN)**	41 (RK,ER)
Cat	(Mouse-Rat)	108 (EE,QN)	
Rat	Golden hamster	**4 (TT,AA)**	

Table S5.6: Convergent and pseudoconvergent sites in GPX2 between the GPX6$_{Cys}$ lineages, where the sequences were also available, as identified by CONVERG2*. The two left-most columns are the branches between which convergent or pseudo-convergent lineages are identified. The number gives the amino acid site where convergence is identified; the brackets represent the (ancestral amino acids on both branches, derived amino acids on both branches) in the order of the branches. Here, (SM-M) stands for the Squirrel monkey – marmoset internal branch, (GH-CH) stands for the Golden hamster – Chinese hamster internal branch and (rat-mouse) represents the rat-mouse internal branch. The branch names given in green indicate the branches upon which we have inferred the Sec-to-Cys exchange in GPX6$_{Cys}$ to have occurred. Strict convergent sites are given in bold; all other sites are pseudo-convergent.*

Rabbit	Mouse	47 (EQ,QE)

***Table S5.7*: Convergent and pseudoconvergent sites in GPX3 between the GPX6$_{Cys}$ lineages, where the sequences were also available, as identified by CONVERG2.** *The two left-most columns are the branches between which convergent or pseudo-convergent lineages are identified. The number gives the amino acid site where convergence is identified; the brackets represent the (ancestral amino acids on both branches, derived amino acids on both branches) in the order of the branches. Here, (SM-M) stands for the Squirrel monkey – marmoset internal branch, (GH-CH) stands for the Golden hamster – Chinese hamster internal branch and (rat-mouse) represents the rat-mouse internal branch. The branch names given in green indicate the branches upon which we have inferred the Sec-to-Cys exchange in GPX6$_{Cys}$ to have occurred. Here, the marmoset branch is given as a branch where the Sec-to-Cys exchange in GPX6$_{Cys}$ has been estimated, given that the squirrel monkey sequence is unavailable for this protein. Strict convergent sites are given in bold; all other sites are pseudo-convergent.*

Marmoset	Eumuroida	**5 (VV,MM)**	172 (SA,LS)
Marmoset	Rabbit	33 (VI,LV)	
Marmoset	Cat	5 (VV,MG)	
Marmoset	(GH-CH)	33 (VI,LV)	
Eumuroida	Rabbit	135 (VV,IM)	
Eumuroida	Cat	5 (VV,MG)	
Rabbit	Cat	**126 (GG,NN)**	**152 (II,VV)**
Rabbit	Chinese hamster	152 (II,VK)	
Rabbit	(GH-CH)	**33 (II,VV)**	**107 (FF,VV)**
Cat	Mouse	154 (II,LV)	
Cat	Chinese hamster	152 (II,VK)	
Cat	(GH-CH)	154 (II,LV)	
Mouse	(GH-CH)	**154 (II,VV)**	

Table S5.8. Convergent and pseudoconvergent sites in GPX4 between the GPX6$_{Cys}$ lineages, where the sequences were also available, as identified by CONVERG2

No convergent sites found.

Table S5.9: Convergent and pseudoconvergent sites in GPX5 between the GPX6$_{Cys}$ lineages, where the sequences were also available, as identified by CONVERG2. *The two left-most columns are the branches between which convergent or pseudo-convergent lineages are identified. The number gives the amino acid site where convergence is identified; the brackets represent (ancestral amino acids on both branches, derived amino acids on both branches)) in the order of the branches. Here, (SM-M) stands for the Squirrel monkey – marmoset internal branch, GH-CH) stands for the Golden hamster – Chinese hamster internal branch and (rat-mouse) represents the rat-mouse internal branch. The branch names given in green indicate the branches upon which we have inferred the Sec-to-Cys exchange in GPX6$_{Cys}$ to have occurred. Strict convergent sites are given in bold; all other sites are pseudo-convergent.*

(SM-M)	Rabbit	140 (RR,QL)		
(SM-M)	Cat	5 (KK,RN)	**140 (RR,QQ)**	
(SM-M)	Walrus	18 (AT,MS)	**29 (QQ,RR)**	
(SM-M)	Squirrel monkey	151 (LV,VE)		
(SM-M)	Marmoset	145 (SL,LI)		
(SM-M)	Mouse	151 (LL,VM)		
(SM-M)	Golden hamster	5 (KK,RT)		
(SM-M)	Chinese hamster	151 (LL,VM)		
(SM-M)	(GH-CH)	18 (AS,MA)	148 (TS,AA)	
Eumuroida	Rabbit	48 (AA,IS)		
Eumuroida	Cat	**130 (DD,NN)**		
Eumuroida	Mouse	130 (DN,ND)		
Eumuroida	Rat	52 (SL,LT)		
Eumuroida	Golden hamster	17 (FL,LF)	48 (AI,IM)	103 (AV,VA)
Eumuroida	(Mouse-rat)	109 (SY,YF)		
Rabbit	Cat	82 (EE,GK)	140 (RR,LQ)	
Rabbit	Squirrel monkey	0 (KQ,QK)		
Rabbit	Rat	**13 (DD,NN)**	86 (KK,NE)	
Rabbit	Golden hamster	0 (KK,QR)	48 (AI,SM)	
Rabbit	(Mouse-rat)	**65 (GG,KK)**	**67 (YY,FF)**	
Cat	Mouse	83 (KK,TN)	130 (DN,ND)	
Cat	Golden hamster	5 (KK,NT)		
Walrus	Marmoset	21 (GK,EE)		
Walrus	(Mouse-rat)	**26 (QQ,PP)**		
Walrus	(GH-CH)	18 (TS,SA)		
Squirrel monkey	Mouse	151 (VL,EM)		
Squirrel monkey	Golden hamster	0 (QK,KR)		
Squirrel monkey	Chinese hamster	151 (VL,EM)		
Marmoset	Chinese hamster	**94 (SS,AA)**		
Mouse	Chinese hamster	**151 (LL,MM)**		
Mouse	(Mouse-rat)	104 (TS,MT)	155 (NK,SN)	
Golden hamster	(GH-CH)	101 (IT,VI)		

Table S5.10: Convergent and pseudoconvergent sites in GPX7 between the GPX6$_{Cys}$ lineages, as identified by CONVERG2. *The two left-most columns are the branches between which convergent or pseudo-convergent lineages are identified. The number gives the amino acid site where convergence is identified; the brackets represent (ancestral amino acids on both branches, derived amino acids on both branches) in the order of the branches. Here, (SM-M) stands for the Squirrel monkey – marmoset internal branch, (GH-CH) stands for the Golden hamster – Chinese hamster internal branch and (rat-mouse) represents the rat-mouse internal branch. The branch names given in blue indicate the branches upon which we have inferred the Sec-to-Cys exchange in GPX6$_{Cys}$ to have occurred. Here, the marmoset branch is given as a branch where the Sec-to-Cys exchange in GPX6$_{Cys}$ has been estimated, given that the squirrel monkey sequence is unavailable for this protein Strict convergent sites are given in bold, all other sites are pseudo-convergent.*

Marmoset	Mouse	95 (AA,SD)		
Marmoset	(GH-CH)	95 (AA,SD)		
Eumuroida	Rabbit	**109 (SS,PP)**		
Eumuroida	Chinese hamster	116 (HR,RQ)		
Rabbit	Mouse	30 (HY,RH)		
Mouse	(GH-CH)	**49 (SS,TT)**	**95 (AA,DD)**	111 (EE,AQ)

Table S5.11: Convergent and pseudoconvergent sites in GPX8 between the GPX6$_{Cys}$ lineages, as identified by CONVERG2. *The two left-most columns are the branches between which convergent or pseudo-convergent lineages are identified. The number gives the amino acid site where convergence is identified; the brackets represent (ancestral amino acids on both branches, derived amino acids on both branches) in the order of the branches. Here, (SM-M) stands for the Squirrel monkey – marmoset internal branch, (GH-CH) stands for the Golden hamster – Chinese hamster internal branch and (rat-mouse) represents the rat-mouse internal branch. The branch names given in blue indicate the branches upon which we have inferred the Sec-to-Cys exchange in GPX6$_{Cys}$ to have occurred. Strict convergent sites are given in bold; all other sites are pseudo-convergent.*

Eumuroida	(GH-CH)	12 (LF,FY)
Rabbit	(GH-CH)	121 (VI,IV)
Mouse	Chinese hamster	59 (KK,QR)
Rat	(Mouse-rat)	18 (QL,EQ)
Chinese hamster	(GH-CH)	51 (MK,TM)

Figures

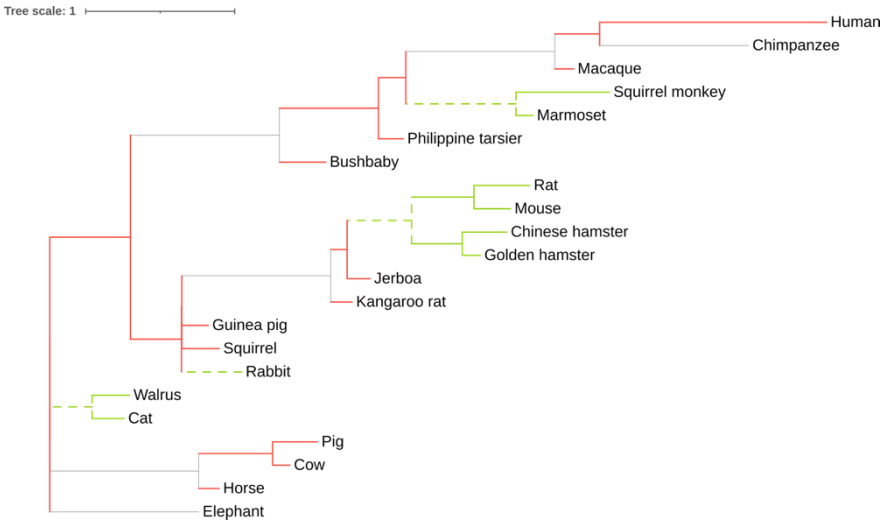


Figure S5.1: Phylogeny of the 22 mammals in our analysis*. In red, GPX6$_{Sec}$ branches, in green, GPX6$_{Cys}$ ones. Branch lengths are proportional to their corresponding dN/dS as estimated by the free-ratio model in PAML (Yang, 2007). In some lineages, PAML estimates very few synonymous changes compared to non-synonymous changes and this results in an unnaturally large dN/dS value. These branches are assigned a dN/dS ratio of 1 and coloured grey, with the actual ratio estimated by PAML is given in parenthesis (#). Branches with dN/dS values given as less than 0.01 are not labelled.*

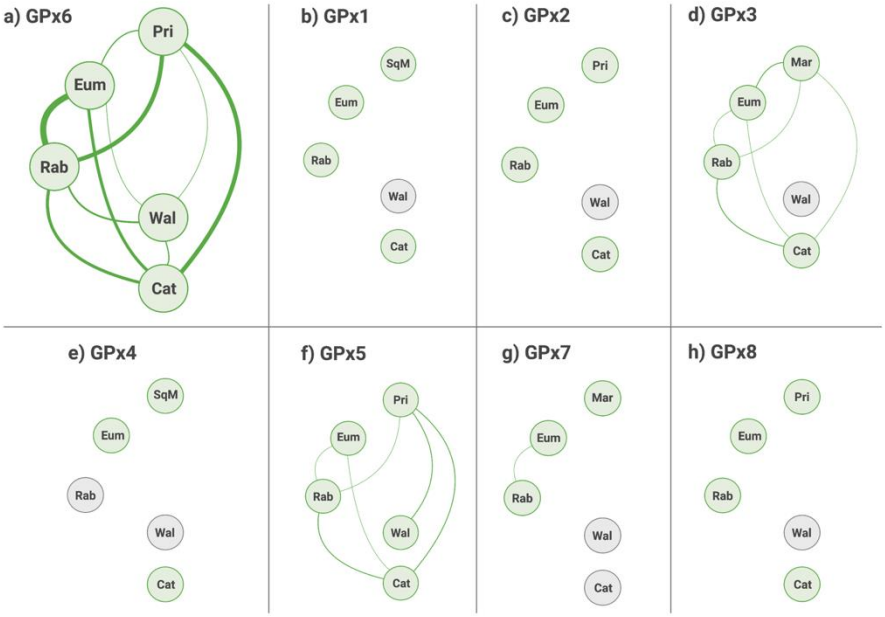


Figure S5.2: Schematic diagram demonstrating the convergence between branches in GPX6 where Sec was inferred to have been lost. *Connection thickness is proportional to the number of convergent sites identified. When a species protein is unavailable, the node is in grey. Pri = primate branch (leading to squirrel monkey and marmoset; Fig 1); Eum= Eumuroida; Rab = Rabbit; Wal = Walrus; Cat = Cat; SqM = Squirrel monkey; Mar = Marmoset*

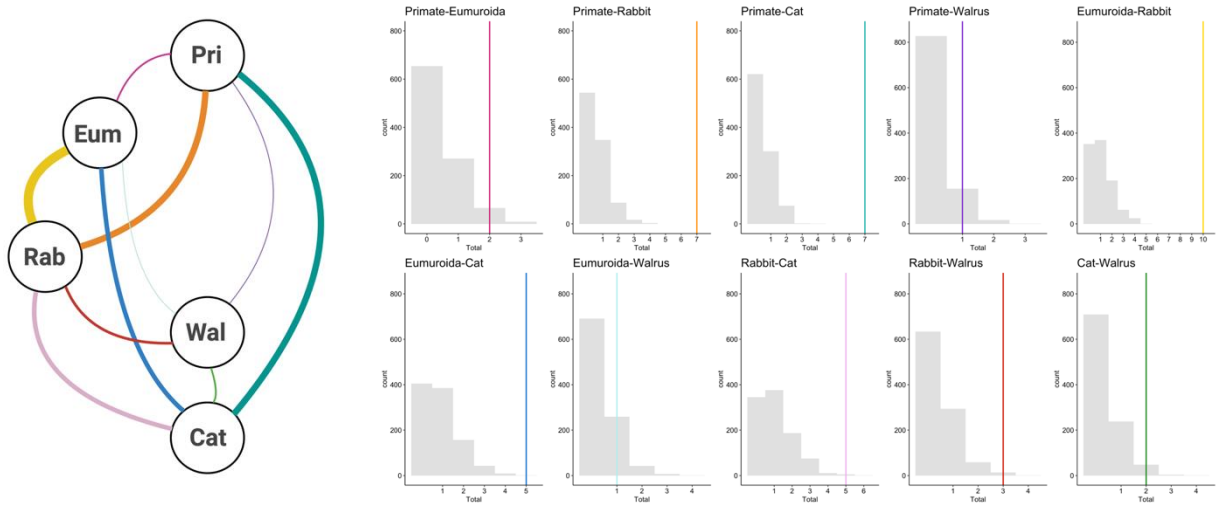


Figure S5.3: Schematic representation of the number of observed convergences in the GPX6 protein between lineages where Sec is lost for Cys*. Thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in the full GPX6 protein between lineages where Sec is lost for Cys according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

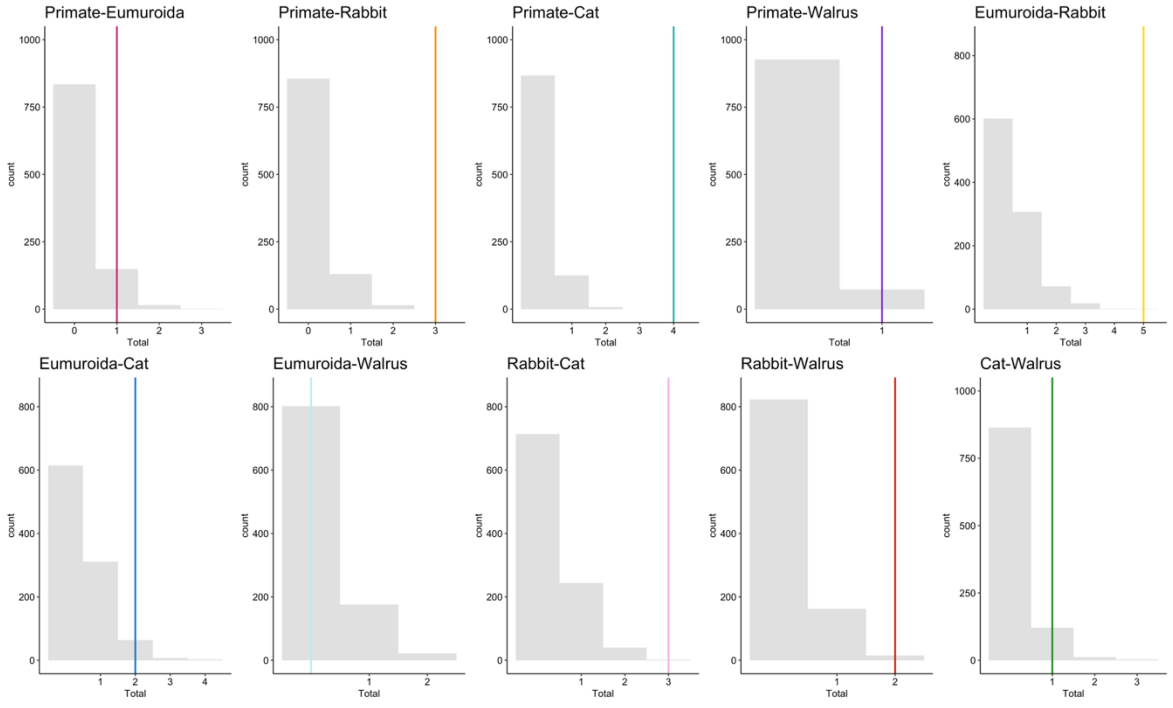


Figure S5.4: Expected distribution of convergent changes in the GPX domain of the GPX6. *Given between lineages where Sec is lost for Cys according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines.*

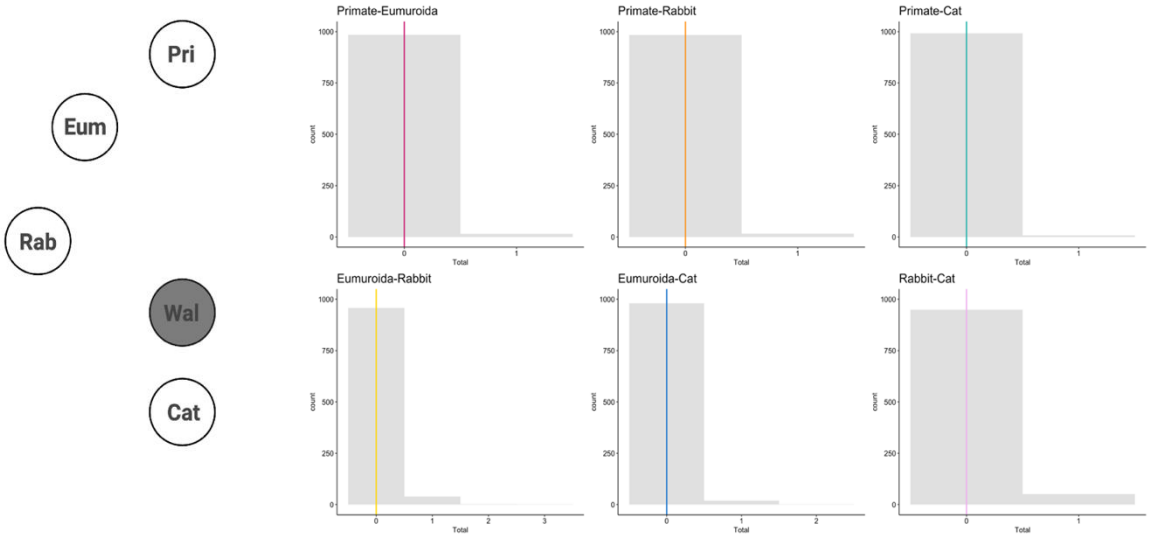


Figure S5.5: Schematic representation of the number of observed convergence in the GPX1 protein*. Given between lineages where Sec is lost for Cys in GPX6, where thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in GPX1 between lineages where Sec is lost for Cys in GPX6 according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

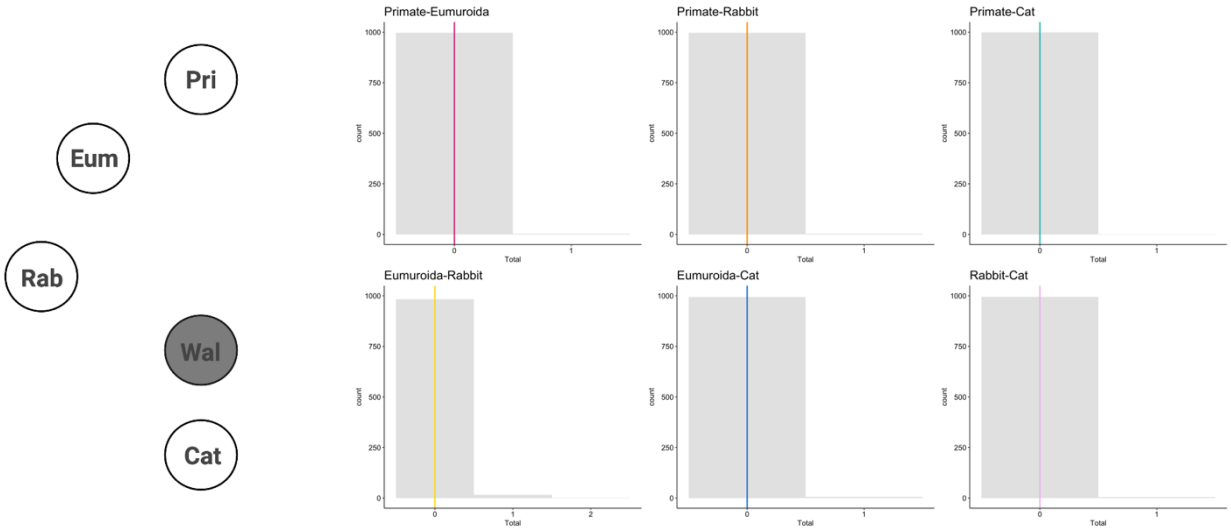


Figure S5.6: Schematic representation of the number of observed convergence in the GPX2 protein*. Given between lineages where Sec is lost for Cys in GPX6, where thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in GPX2 between lineages where Sec is lost for Cys in GPX6 according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

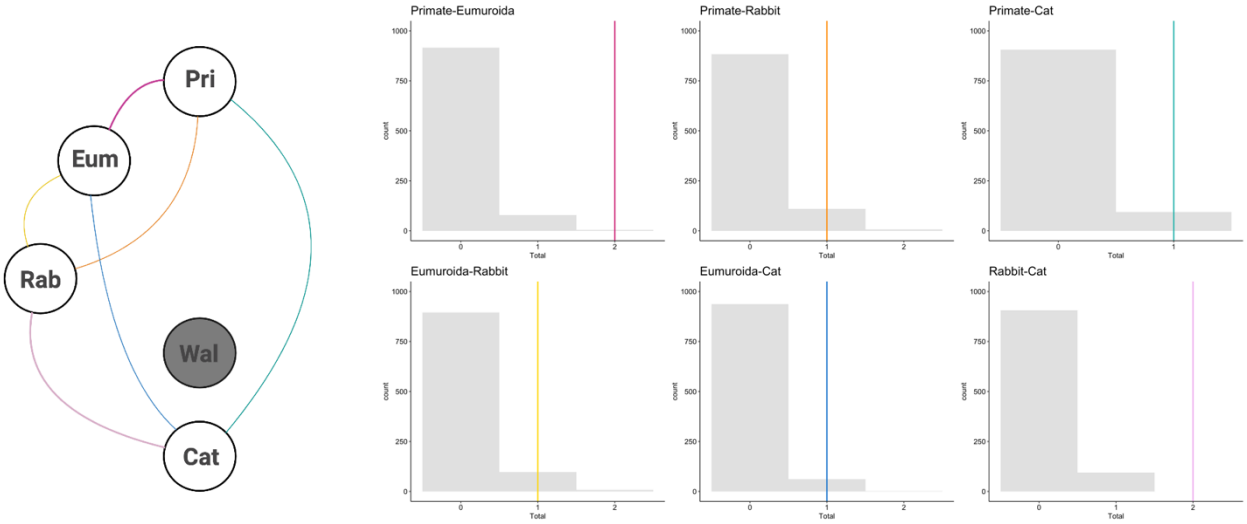


Figure S5.7: Schematic representation of the number of observed convergence in the GPX3 protein. *Given between lineages where Sec is lost for Cys in GPX6, where thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in GPX3 between lineages where Sec is lost for Cys in GPX6 according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

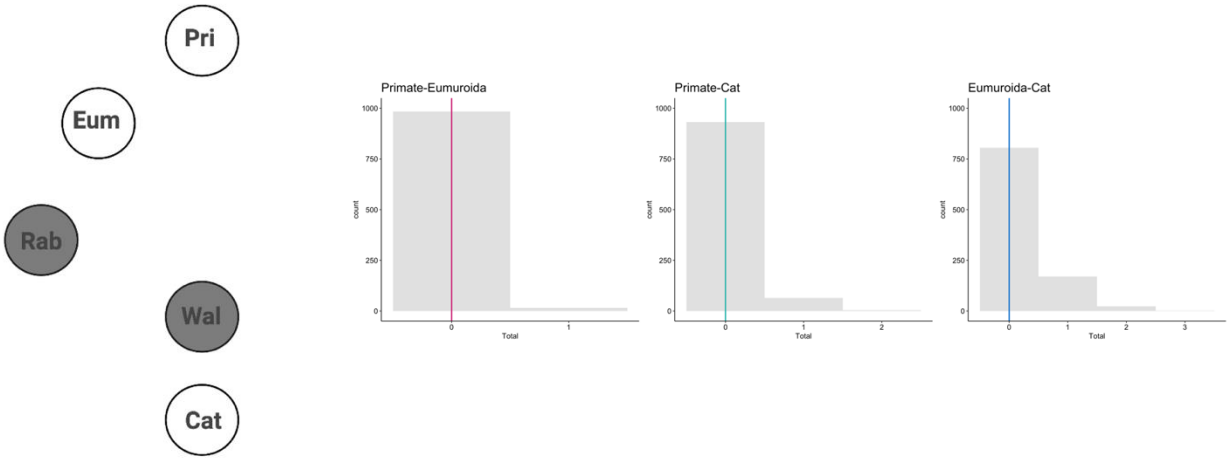


Figure S5.8: Schematic representation of the number of observed convergence in the GPX4 protein*. Given between lineages where Sec is lost for Cys in GPX6, where thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in GPX4 between lineages where Sec is lost for Cys in GPX6 according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

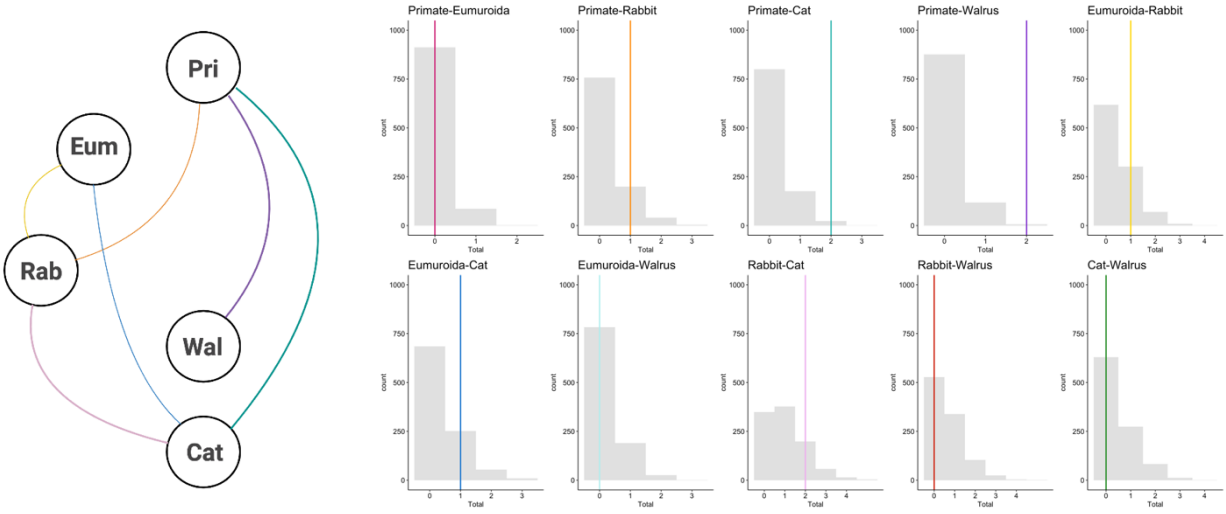


Figure S5.9: Schematic representation of the number of observed convergence in the GPX5 protein*. Given between lineages where Sec is lost for Cys in GPX6, where thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in GPX5 between lineages where Sec is lost for Cys in GPX6 according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

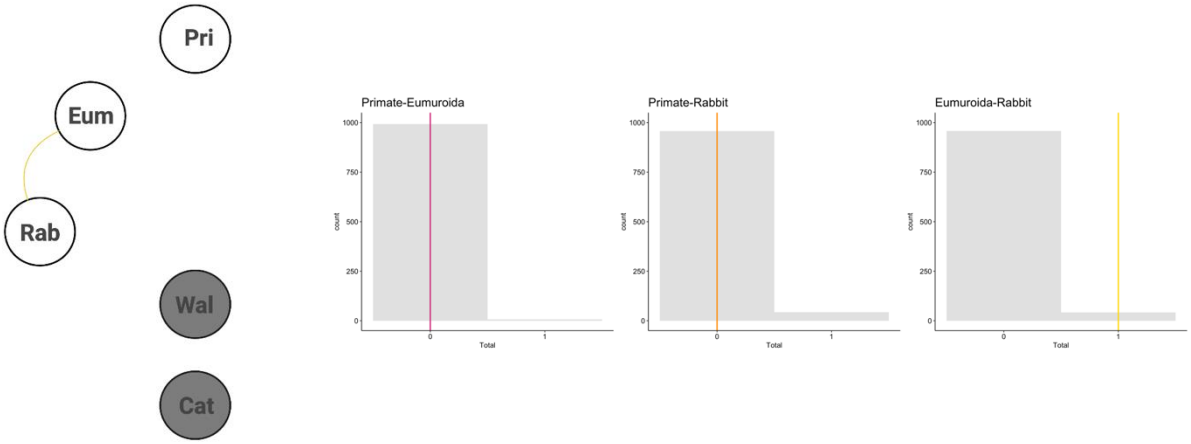


Figure S5.10: Schematic representation of the number of observed convergence in the GPX7 protein. *Given between lineages where Sec is lost for Cys in GPX6, where thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in GPX7 between lineages where Sec is lost for Cys in GPX6 according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

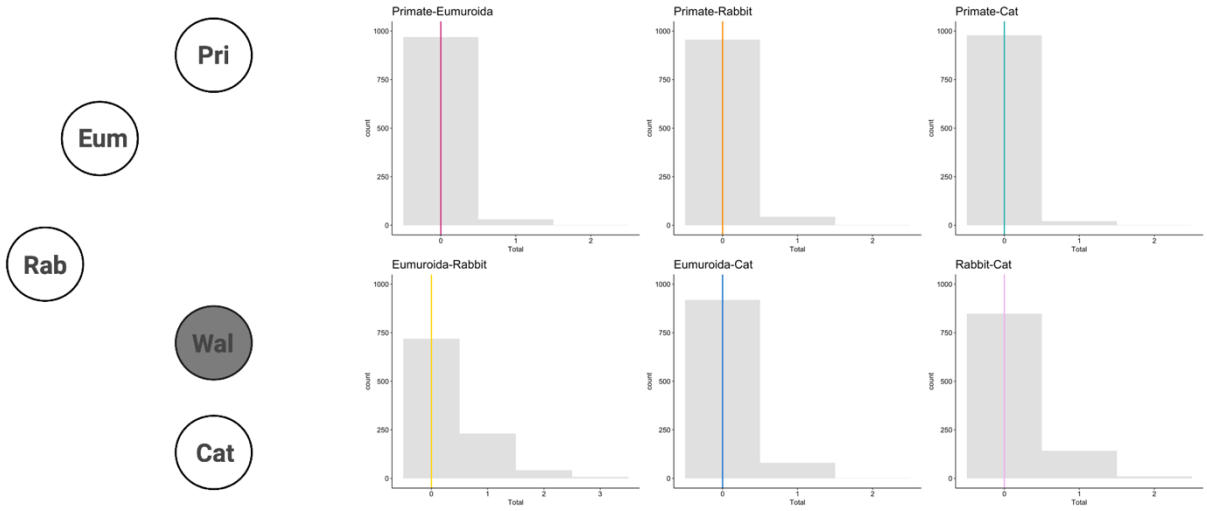


Figure S5.11: Schematic representation of the number of observed convergence in the GPX8 protein. *Given between lineages where Sec is lost for Cys in GPX6, where thickness of the line represents the number of convergent changes (left). Expected distribution of convergent changes in GPX8 between lineages where Sec is lost for Cys in GPX6 according to our Seq-Gen simulations, where the observed numbers of convergent changes are given by coloured lines (right).*

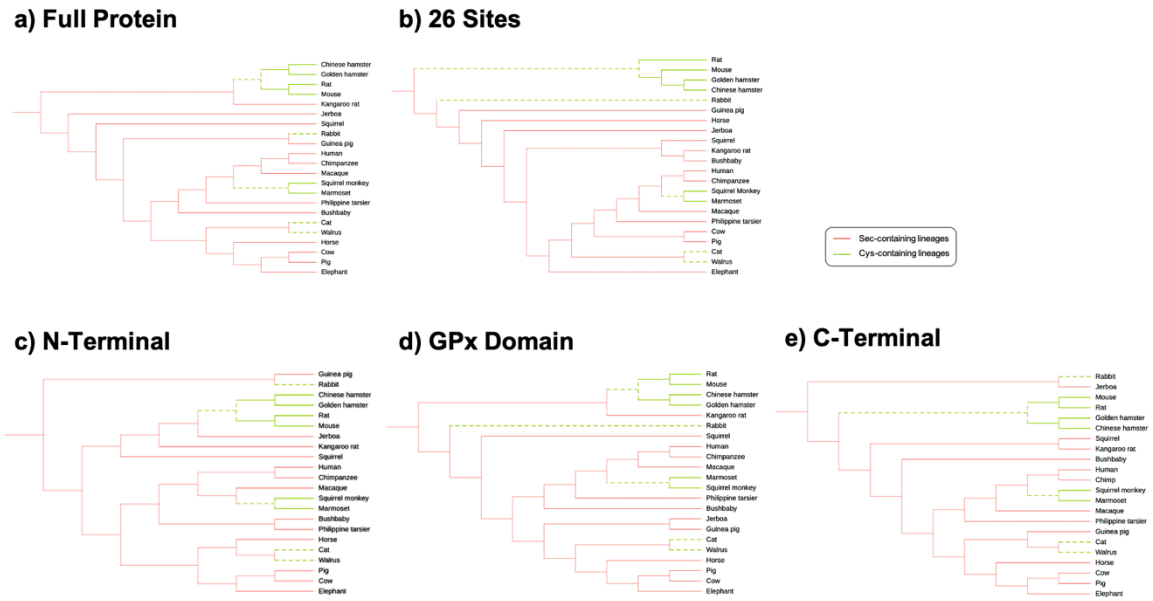


Figure S5.12: Topology of the phylogenetic tree for GPX6, with midpoint rooting, constructed using PHYML. *Shown for the* ***a****) full GPX6 protein;* ***b****) the 26 sites that differ between Eu-GPX6$_{Sec}$ and Eu-GPX6$_{Cys+25}$;* ***c****) the N-terminal of GPX6;* ***d****) the GPX domain of GPX6 and* ***e****) the C-terminal of GPX6. In red, GPX6$_{Sec}$ branches, in green, GPX6$_{Cys}$ ones. Dashed green branches represent GPX6$_{Cys}$ lineages at the time Sec was lost.*

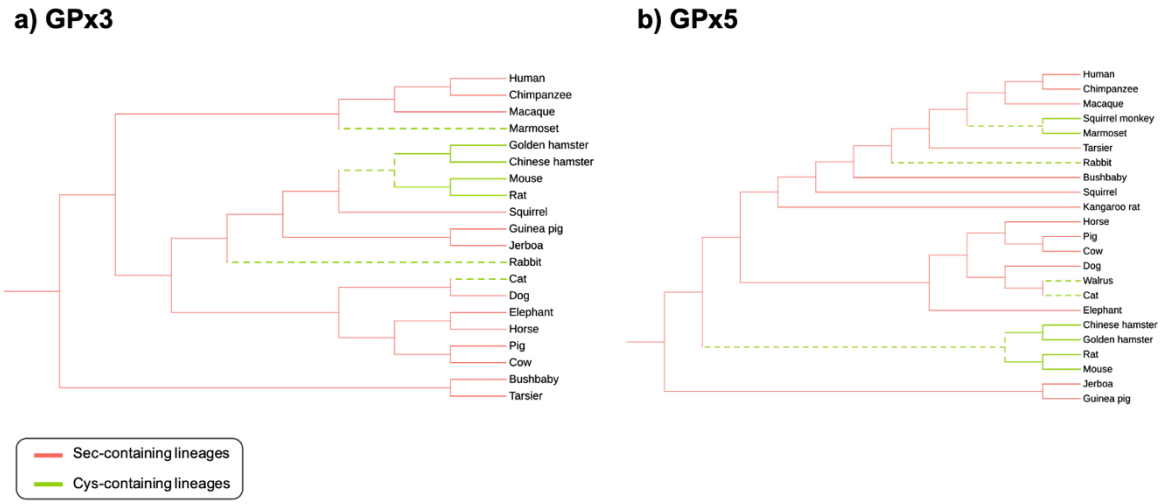
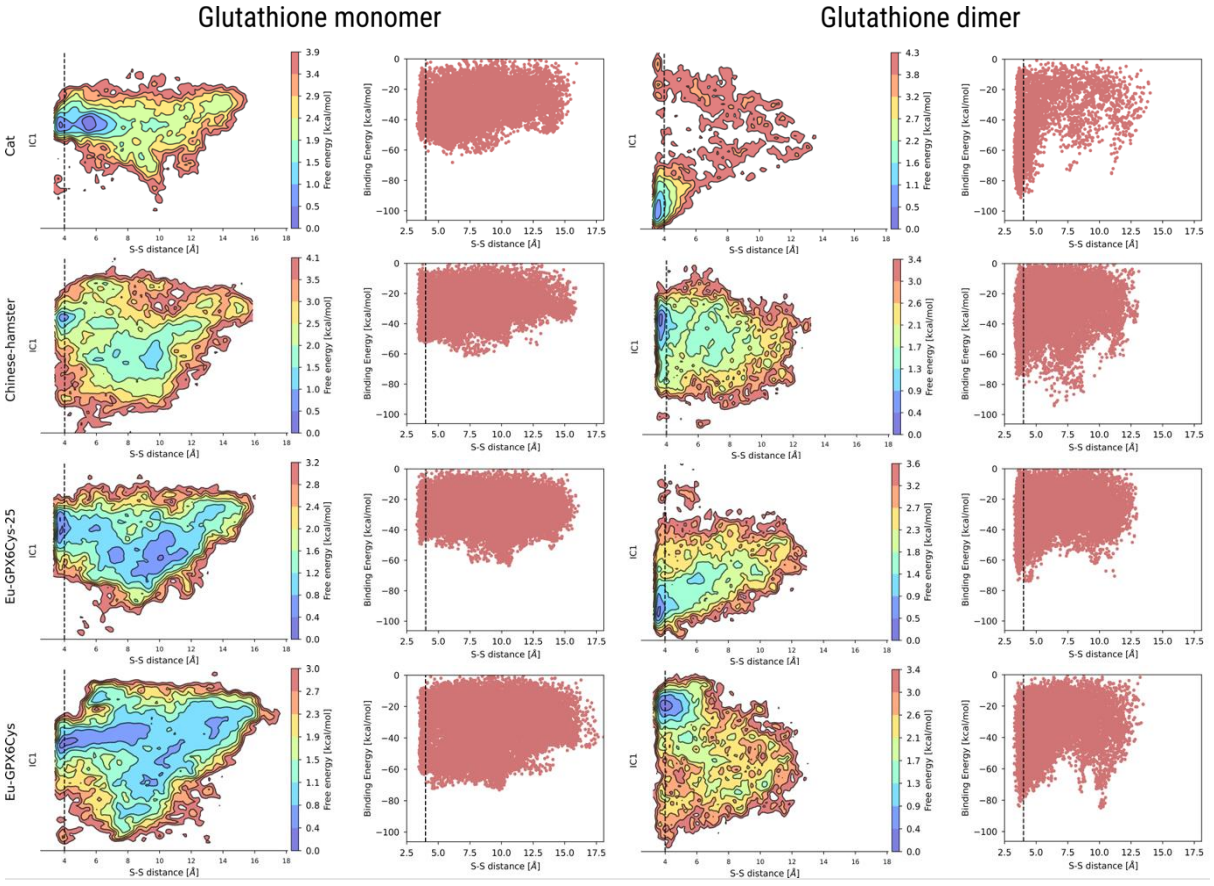


Figure S5.13: Topology of the phylogenetic trees, with midpoint rooting, constructed using PHYML for additional GPX proteins*. Shown from the available mammalian proteins of a) GPX3 and b) GPX5. In green, $GPX6_{Cys}$ ones. Dashed green branches represent $GPX6_{Cys}$ lineages at the time Sec was lost.*



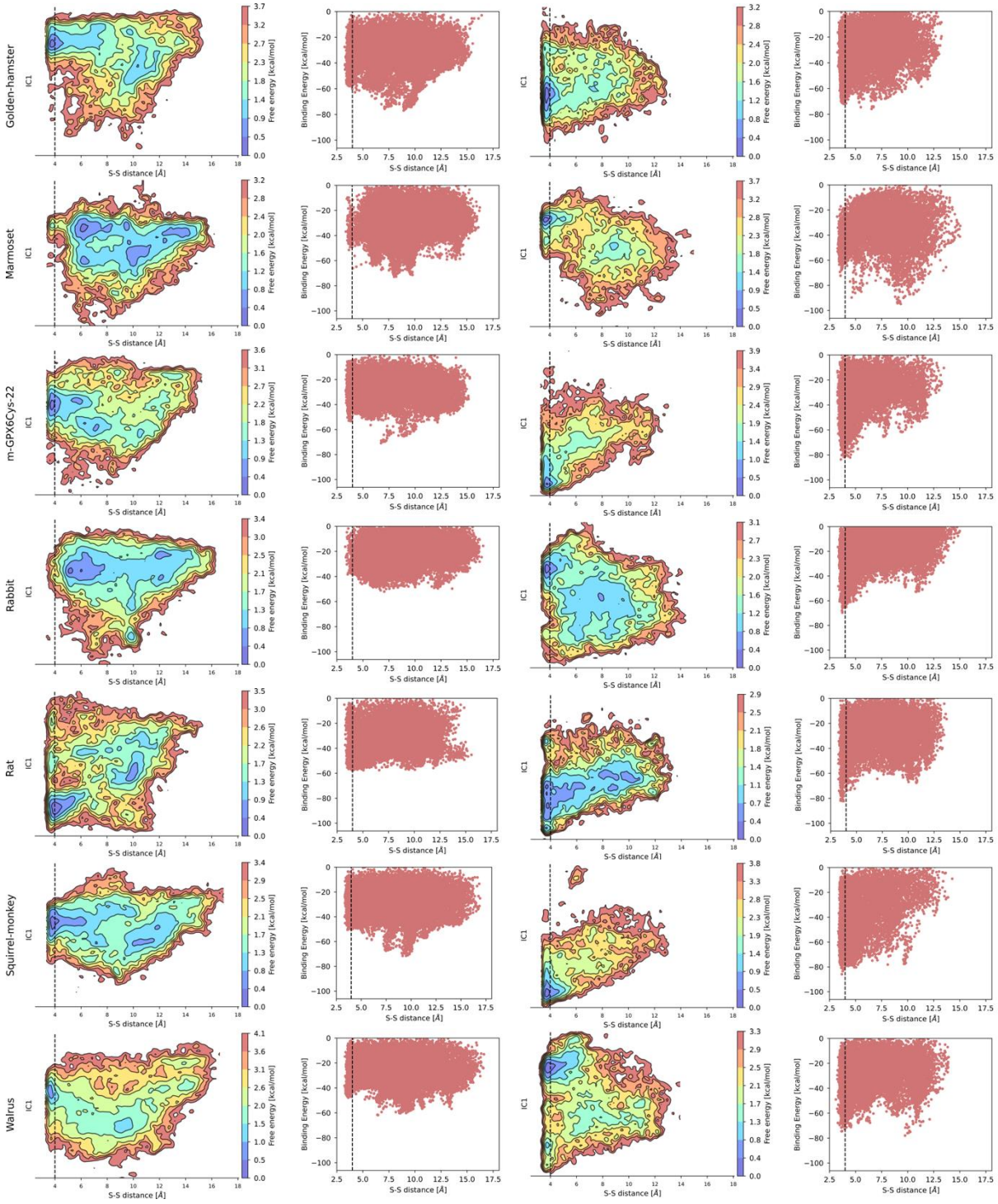


***Figure S5.14**: **Free energy profiles for the docking of glutathione (left) and glutathione disulfide (right) to ancestral and modern GPX6$_{Cys}$ proteins.** The x-axis represents the distance between the catalytic cysteine sulphur atom and the closest ligand's sulphur atom, while the Y-axis shows the slowest TICA coordinate or the binding free energy. The vertical dashed line represents a distance of 4 Å.*