

The genomics of selenium: its past, present and future

Tom Davy and Sergi Castellano^{1,2}

¹Genetics and Genomic Medicine Programme
Institute of Child Health (ICH)
University College London (UCL)
London, UK

²UCL Genomics
University College London (UCL)
London, UK

Abstract

In the last two decades, genomic analyses have enriched the study of the biology of selenium in many ways. These include the identification of selenoproteins in prokaryotic and eukaryotic genomes, the discovery of genetic variants that mediate humans and other vertebrates' adaptations to their selenium nutritional histories, and the association of specific genotypes with common and rare human selenium disorders. We briefly review these computational, evolutionary and association studies and their contribution to the genomics of selenium, selenocysteine and selenoproteins in the 200th anniversary of the discovery of this trace element.

Keywords: selenium, genomics, gene prediction, adaptation, common disease, rare disease

Introduction

Selenium is an essential micronutrient, for which consumption levels in humans and other species are closely tied to local soil levels. This trace element is utilised most notably in the manufacture of proteins wherein a selenocysteine (sec, U) residue is found in place of cysteine. Selenocysteine is encoded for by the UGA stop codon and translated via a specifically conserved suite of RNA and protein machineries. Such selenoproteins number 25 in humans, with varying figures throughout species across the three life's domains ranging from zero to 59 (Lobanov, et al. 2009). A myriad of species employs selenoproteins to harness the reductive and catalytic capabilities of selenium to counter oxidative stress and, possibly, perform other less common functions (*e.g.* a structural role for glutathione peroxidase 4 and perhaps for SELENOJ) (Ursini, et al. 1999; Castellano, et al. 2005).

This review will address the history of selenoprotein and selenium-associated protein research through genomics, with a focus on the computational identification of selenoprotein genes in genomes, the genomic signatures of adaptation to pressures incited by changes in environmental selenium content and the role of modern population genomics to study human disease variants related to selenium and selenoproteins. The increased genetic resolution (from genes to genotypes) brought by genomics to selenium research is discussed throughout.

1. Entering the genomic era

Though it was discovered by Jons Jacob Berzelius in 1817, it would take a further 140 years before selenium was recognized as an essential micronutrient to both humans and other species, overcoming the prevailing idea that it was only a toxic substance. What's more, it would take until 1976 before the presence of the element was confirmed to exist within the sequence of proteins (Cone, et al. 1976). Further analyses revealed a replacement of cysteine with the serine-derived selenocysteine (Forstrom, et al. 1978) for certain cysteine sites in particular proteins. More so unique was the identity of a specifically responsible codon, appraised from the cloning of the genes glutathione peroxidase 1 and formate dehydrogenase, discovered to be TGA (or UGA in the mRNA) (Chambers, et al. 1986; Zinoni, et al. 1986), also known as the "opal" stop codon, cementing the position of selenocysteine as the 21st amino acid in the genetic code (Forchhammer and Bock 1991).

This groundwork proved fundamental to the understanding of selenoprotein biochemistry, molecular biology and genetics. Unique mechanics for the insertion of selenocysteine began to be uncovered with the identification of a unique Sec-tRNA alongside accompanying upstream enzymes (Forchhammer and Bock 1991). A high degree of conservation meanwhile suggested an ancient origin for selenoproteins (Guimaraes, et al. 1996; Stadtman 1996).

Analysis of the mRNA sequences in search for a selenocysteine insertion mechanism lead to the discovery of the SECIS (SElenoCysteine Insertion Sequence) element in prokaryotes and eukaryotes (Zinoni, et al. 1990; Berry, et al. 1991) (Liu, et al. 1998) through acknowledgment of a conserved secondary structure, accompanied by site-directed mutagenesis (Walczak, et al. 1996; Berry, et al. 1997).

1.1. Transcript annotation

Laboratory bench-based screenings had been successful at identifying multiple novel selenoproteins by the late 1990s. Indeed, twelve mammalian selenoproteins, from seven protein families, had been discovered by 1999 using various techniques (Holben and Smith 1999), and homologs and other related genes were being identified throughout the three domains of life (Persson, et al. 1997; Wilting, et al. 1997; Zhong, et al. 1998; Gladyshev and Hatfield 1999). However, as increasing amounts of genetic data were gathered, swifter computational tools promised a more comprehensive description of the selenoproteome in different organisms.

Genetic sequences, such as those from the NCBI's mRNA expression sequence tag database (dbEST) (Tugendreich, et al. 1993) were becoming approachable by computational methods that aimed to identify selenoproteins from previously identified sequence motifs. The end of 1999 saw two papers applying varying computational methods sequences to successfully identify selenoproteins both *de-novo* and *in-silico* simultaneously. Lescure *et al.*, 1999, leveraged the dbEST database to find three new mammalian selenoproteins (SelX, SelN and SelZ, with two spliceforms SelZf1 and SelZf2, in their original nomenclature) via an RNA secondary-structure prediction methodology, wherein hairpin loops were identified (Lescure, et al. 1999). Meanwhile Kryukov *et al.*, developed SECISearch, a program that searched for thermodynamically stable SECIS motifs within the 3' UTR of mRNA sequences, discovering two new selenoproteins dubbed SelT and SelR, although SelX is equivalent to SelR and SelZ to SelT using their original names (Kryukov, et al. 1999). SelR is now known as MSRB1 and SelT as SELENOT (Gladyshev, et al. 2016). Such searches were not limited to cellular life, with even viral databases being subject to similar approaches, identifying a potential GPX in *Molluscum contagium* using similar comparative techniques (Zhang, et al. 1999). The accelerating pace of DNA sequencing promised complete genomes however, and as such new methods were needed to identify selenoprotein genes in long genomic sequences.

1.2. Genome annotation

The first instances of the use of genomic sequences to characterise an organism's complete selenoproteome were Martin-Romero *et al.*, 2001 and Castellano *et al.*, 2001. In the first of this works, the annotations of standard genes in the *Drosophila melanogaster* genome were

coordinated with the downstream presence of SECIS elements (Martin-Romero, et al. 2001). Two novel selenoproteins were uncovered (albeit they were called G-rich and BthD, respectively). In the second work, published in the same month, SECIS element RNA secondary structure prediction was utilised parallel to the *ab initio* prediction of genes wherein a sequence possessed a strong codon-bias past a predicted UGA selenocysteine codon, essentially capturing misannotated stop codons along the genome (Castellano, et al. 2001). A selenoproteome size of just three proteins was predicted for *Drosophila melanogaster*, with the same two predictions being novel selenoproteins (SelG and SelH in the original designation). These selenoproteins were shown to incorporate radioactively-labelled selenium, demonstrating the limitations of standard gene annotations for selenoproteins in eukaryotic genomes. SelG is known as SELENOK and SelH as SELENOH (Gladyshev, et al. 2016).

The same methods were adopted to annotate the mice, rat and human selenoproteomes, alongside checking for conservation of selenocysteine or cysteine at predicted positions for additional validation, expanding the human selenoproteome from 18 to 25 proteins (Kryukov, et al. 2003). In contrast, comparative gene prediction approaches that made use of two genomes, irrespective of their SECIS elements, were used to identify SelU in the fugu genome, a puffer fish, which contains cysteine in mammals (Castellano, et al. 2004). One worrying limitation of such comparative approaches is that it required the presence of Sec or Cys-containing orthologs in the compared species. Indeed, the *ab initio* prediction of selenoprotein genes in the Tetraodon genome (another puffer fish) was necessary to identify the SelJ gene, which has no counterpart in mammals (Castellano, et al. 2005). These and other analyses (Lobanov, et al. 2007; Mariotti, et al. 2012; Mariotti, et al. 2015) showed that the size of the selenoproteome varies widely across vertebrates, with fishes having from 10 to 14 more selenoprotein genes than other vertebrate species. Such variance was mirrored across non-vertebrates with the confirmation of the *Caenorhabditis elegans* selenoproteome to be just one protein (Taskov, et al. 2005), with other nematode species having none (Otero, et al. 2014) or a few more (Taskov, et al. 2005). *Drosophila* species have none (Chapple and Guigo 2008) or up to three selenoproteins (Martin-Romero, et al. 2001) (Castellano, et al. 2001). Interestingly, land plants and fungi (including the single-cell yeasts) have no selenoproteins and use cysteine-containing proteins instead (Lobanov, et al. 2009). In contrast, the green algae *Chlamydomonas reinhardtii* has 10 selenoproteins with other unicellular algae having as many as 59, the largest eukaryotic selenoproteome to date (Gobler, et al. 2013). Other eukaryotic species have variable number of selenoproteins (Lobanov, et al. 2009).

Prokaryote annotation also proved challenging at times (Lobanov, et al. 2006), with specialist methodologies being adopted for the annotation of prokaryotic selenoproteomes (Kryukov and Gladyshev 2004; Zhang and Gladyshev 2005). Most of these new selenoproteins were identified by taking advantage of the fact that selenocysteine and cysteine are often found in homologous

positions in proteins from different species (Zhang, et al. 2005), albeit methods to identify prokaryotic SECIS elements also proved fruitful (Zhang and Gladyshev 2005). The surge in the environmental sequencing of prokaryotic genomes allowed these approaches to identify hundreds of additional selenoproteins (Zhang, et al. 2005; Zhang and Gladyshev 2008), thus contributing to today's thorough description of the prokaryotic use of selenium in proteins (Peng, et al. 2016). The largest selenoproteome in prokaryotes stands so far at 57 selenoproteins, in a symbiotic bacteria (Zhang and Gladyshev 2007).

Many selenoproteomes are now annotated and some of them are available in SelenoDB (Romagne, et al. 2014) or other specialized databases (Bekaert, et al. 2010). Most animal selenoproteomes were originally annotated manually (Castellano, et al. 2008) but automatic methods to annotate selenoproteins have now been developed (Mariotti and Guigo 2010; Mariotti, et al. 2013). These methods have become increasingly important as the pace of genome sequencing has kept increasing (Mariotti 2018), making manual annotation impractical.

1.3. Adaptation in vertebrates

The annotation of selenoproteins in multiple genomes has opened new avenues for evolutionary enquire. For example, they have allowed to pose the question of whether natural selection has distinctly shaped the use and regulation of selenium among the different vertebrate clades. Levels of selenium vary hundreds-fold across lands and waters worldwide, but aquatic environments collect land selenium (Selinus, et al. 2005), with its inorganic forms being efficiently accumulated in phytoplankton and converted into organic forms that enter the animal diet. Vertebrates from different environments have thus evolved with different dietary amounts of selenium. This is important because selenium deficiency, in particular, is a potential selective factor in vertebrate evolution. One reason for this is that the use of selenocysteine is conserved (under purifying selection) in vertebrates (Castellano, et al. 2009), suggesting that selenocysteine is not easily replaced in proteins by cysteine. As a result, vertebrates rely on selenium even when this is scarce (or toxic). A second reason is that the metabolism and homeostasis of selenium in vertebrates has regulatory features seemingly compatible with adaptation to selenium scarcity. One of this is the hierarchical prioritization of its supply to certain tissues and selenoproteins in the absence of sufficient selenium in the diet (Sunde and Raines 2011). Another is the very low interpersonal differences of selenium in liver samples compared to other trace elements (Engelken, et al. 2016), suggesting strong selection for keeping this homeostatic system tightly in check, which is agreeable with the context that the adequate and toxic doses are relatively close (Wilber 1980).

The question above can be addressed comparing the evolutionary forces acting on the coding sequences of selenoprotein genes and genes involved in the regulation of selenium between clades and to those acting on their paralogs with cysteine along the vertebrate phylogeny (Sarangi, et al.

2018). In doing so, it can be shown that the strength of natural selection has significantly changed across vertebrate clades for genes that use or regulate selenium, while it is more constant in the cysteine-containing genes, which use sulfur instead of selenium. Among the different vertebrate clades, the strength of selection is unusually variable in the selenoprotein genes of teleost fishes, suggesting a complex nutritional history in a clade where selenoprotein gene duplications abound. In agreement, teleost fishes subfunctionalized or neofunctionalized selenoprotein genes and maintained their capacity for selenium transport in the body (*via* SELENOP), which neutrally declined for millions of years in terrestrial vertebrates.

These results are interesting as it has been previously suggested that teleost fishes have developed greater dependence on environmental selenium, whereas other vertebrates have reduced their reliance on it (Lobanov, et al. 2007; Lobanov, et al. 2008). Sequence patterns in selenium-related genes in vertebrates support this hypothesis, with the essentiality of selenium in vertebrate proteins and its environmental variation across lands and waters being a selective pressure for genes and proteins that use or regulate this trace element, with both its deficiency in humans (White, et al. 2015) and other terrestrial vertebrates and abundance in teleost fishes having left signatures of natural selection. Whether this is also the case in invertebrates remains unknown. Invertebrates are particularly interesting from an adaptive point of view as a number of species have independently lost the use of selenocysteine and employ cysteine in their proteins instead (Chapple and Guigo 2008; Otero, et al. 2014; Sadd, et al. 2015). Still, whether adaptive or neutral processes are responsible for the loss of Sec in the different invertebrate lineages remains unclear, and further evolutionary tests on the sequence patterns of each lineage are needed.

2. Entering the population era

The sequencing of thousands of human exomes and genomes from populations around the world has revolutionized the study of human history (Genomes Project, et al. 2015), including the history of human adaptations to their local environment. Such adaptations have enabled humans to settle in the many distinct environments across the earth (Sabeti, et al. 2007), environments that provide different levels of the micronutrients essential to the human and other species diet. Of these essential micronutrients, selenium is atypical due to its small margin between nutritionally optimal and potentially toxic (Wilber 1980), making its uneven environmental distribution challenging. Diet is the main source of selenium and its intake depends on the its content in the soil on which food is gathered, hunted or grown. Soil selenium levels depend largely on the underlying bedrock from which they are formed, which has resulted in a patchwork of deficient, adequate and sometimes toxic areas across the world varying hundreds-fold in their selenium levels.

2.1. Adaptation in humans

Diets around the world vary in their composition, and this variance provides scope for adaptation to the local environment. Examples of nutrition-related adaptations include the ability to digest the carbohydrate lactose beyond infancy due to changes in the expression of the lactase enzyme in different dairy herding populations (Tishkoff, et al. 2007), the improved digestion of starch in populations that have historically relied on agricultural diets (Perry, et al. 2007), and the capacity to cope with large amounts of polyunsaturated fatty acids in the Greenland Inuit due to their specialized animal fat-rich diet (Fumagalli, et al. 2015). These local adaptations are the result of allele frequency changes on common variation or new mutations. Adaptations to micronutrient deficiencies have also been proposed. For example, to iron deficiency in Europe where an iron accumulation phenotype (homozygous mutations in the HFE iron transporter) may have become advantageous with the introduction of farming and the subsequent lower iron content in the diet (Beutler 2006). Similarly, some African populations living in iodine poor environments may have adapted to its deficiency (Lopez Herraiez, et al. 2009).

Human genetic variation in selenoprotein genes and other genes involved in the metabolism and homeostasis of selenium has not been investigated on a global scale until recently, albeit patterns of genetic variation in glutathione peroxidase 1 may be compatible with positive selection in Asia (Foster, et al. 2006). To examine the possibility of local adaptation driving selenoprotein evolution, coding and non-coding polymorphisms have to be uniformly ascertained across the world. Such variation data has become available for 53 human populations (Romagne, et al. 2014). These are ethnic populations from Africa, the Middle East, Europe, Asia, Oceania and America with distinct selenium nutritional histories. These populations have not experienced recent migrations, making them suitable for studies of local adaptation (Cann, et al. 2002). Thus, selenium-related genes with unusually high levels of differentiation between populations are candidates for having experienced adaptation to local selenium levels. Interestingly, some selenoproteins genes and genes that regulate selenium exhibit, as a group, levels of differentiation beyond what is expected under neutral evolution in populations likely to be under selenium deficiency (White, et al. 2015). These populations inhabit areas of China where Keshan cardiomyopathy and Kashin-Beck osteoarthropathy were endemic before selenium supplementation. This suggests that polygenic adaptation has occurred recently among human populations in response to selenium deficiency. The functional consequences of the genetic variants behind these adaptations are however unclear but the population frequencies and function of some, when known, are seemingly compatible with adaptation to low levels of dietary micronutrients (White, et al. 2015). In contrast, genes that contain cysteine and are paralogous to selenoprotein genes do not show signatures of local adaptation in these Chinese populations. Thus, selenium appears to have been important in recent

human evolution and may have helped some populations to inhabit environments that are deficient in this element. It will be interesting to investigate population sequence data from other vertebrate species to assess recent genetic adaptations to dietary selenium, particularly since humans started to domesticate animals thousands of years ago (Driscoll, et al. 2009).

3. Entering the patient era

The study of thousands of healthy individuals from different human populations has allowed to understand the functional consequences of certain genotypes in selenoprotein function and regulation (Gudmundsson, et al. 2009; Geller, et al. 2014; Perry, et al. 2014; Zhang, et al. 2017). More recently, the sequencing of diseased individuals has provided insights into selenoprotein disease biology, with inferences now being made between clinical phenotypes and specific genotypes (Dumitrescu, et al. 2005; Agamy, et al. 2010; Schoenmakers, et al. 2010; Anttonen, et al. 2015; Schoenmakers, et al. 2016). Taken together, the study of interindividual human variation promises to advance the study of selenium, selenocysteine and selenoprotein function and metabolism. This variation includes single nucleotide variants, insertions and deletions, copy number variants and others, both common and rare.

3.1. Common variants

Genetic variants between individual genomes are often shared among many other genomes in a population. This common variation was initially described in a seminal study that provided a catalog of single nucleotide variants in human populations and the genetic association among them (International HapMap 2005). This association, due to linkage disequilibrium, reflects the coinheritance of groups of single nucleotide variants (haplotypes) and their functional annotation is needed to investigate, often using genome-wide association studies, the hereditary factors linked to the complexity of disease. These association studies have, for example, linked common variants in selenium-related genes with gestational duration (Zhang, et al. 2017), risk factors for multiple cancers (Ratnasinghe, et al. 2000; Meplan and Hesketh 2012; Meplan, et al. 2012), diseases of the heart (Alanne, et al. 2007), thyroid disease (Santos, et al. 2014) and inflammatory response (Curran, et al. 2005). Understanding common variation may also help understand selenium metabolism, with some variants contributing towards bodily selenium levels (Gentschew, et al. 2012; Cornelis, et al. 2015).

The study of common variation is also important because signatures of adaptation to selenium deficiency in Chinese populations are the result of subtle frequency shifts in variants that are

common, which are currently difficult to interpret functionally. Still, a few of these variants have known functional consequences. For example, iodothyronine deiodinase 2, an oxidoreductase that catalyses the conversion of the iodine-dependent hormone T4 to its active form T3 in the thyroid, has a Threonine to Alanine substitution with high population differentiation in East Asian populations. Individuals homozygous for the Ala allele exhibit significantly lower enzyme activity (Canani, et al. 2005). The activity of this enzyme relies on two micronutrients, selenium and iodine, whose deficiency has been associated with Kashin–Beck disease risk in China (Yao, et al. 2011). On the other hand, glutathione peroxidase 1, an antioxidant enzyme that reduces hydrogen peroxide thereby protecting cells from oxidative damage, has a Proline to Leucine amino acid change and a noncoding A to G variant in the promoter region (Hamanishi, et al. 2004) with high population differentiation also in East Asian populations. The Leucine variant is found at lower frequencies in East Asia than in the rest of the world (excluding America), results in 40% lower activity than the ancestral Proline variant (Hamanishi, et al. 2004). In addition, the derived G variant decreases 25% the transcriptional activity of the gene and is again found at lower frequencies in East Asia than elsewhere (other than in America). The presence of the ancestral variant conferring higher transcriptional and enzymatic activity could reflect adaptation to selenium deficiency. Other variants in glutathione peroxidase 1 may result in differences in subcellular distribution (Bera, et al. 2014) and changes in the concentrations of proteins which are linked to clinically relevant phenotypes (Ekoue, et al. 2017). A combination of association studies and explicit variant assays will describe the phenotypic impacts of common selenium-related variants.

3.2. Rare variants

Rare variants in humans, which are population-specific and found at low frequency, are to a large extent the result of the recent growth of human populations (Keinan and Clark 2012; Tennessen, et al. 2012). Indeed, humans have accumulated in the last 5,000-10,000 years a large number of rare variants, mostly in non-African populations (Fu, et al. 2013). Protein-coding variants found at low frequency in human populations are often slightly deleterious as inferred by approaches that measure the impact of amino acid changes on the structure or function of proteins. Among these, there are those linked to diseases that are both rare and severe. Rare diseases, understood as those that occur in one out of 2,000 individuals, have often a genetic basis and the annotation of the deleteriousness of the causal mutation(s) in single-gene disorders is necessary for their diagnosis and treatment. There are reasons to think that the sequencing of thousands of diseased individuals will uncover new rare mutations that impact selenium-related genes, as has been already shown for a number of them (Schweizer and Fradejas-Villar 2016). Rare and novel mutations in Selenoprotein N are being clinically identified as contributing towards myopathy, a diagnosable disease characterized by neuromuscular weakness, spinal rigidity and early onset respiratory

deficiency, with increasing frequency due to many individual genomes being sequenced (Ardissone, et al. 2016), while a number of rare mutation in the biosynthesis and incorporation of selenocysteine into proteins are responsible for a spectrum of systemic disorders due to the reduced activity of selenoproteins in the body (Dumitrescu, et al. 2005; Agamy, et al. 2010; Schoenmakers, et al. 2010; Anttonen, et al. 2015; Schoenmakers, et al. 2016).

Related to selenium in the diet, Kashin-Beck is a notable disease with a genetic component, with a large body of evidence linking susceptibility to this disease with both population-specific and selenium-related gene variants (Du, et al. 2015; Xiong, et al. 2015) alongside areas of selenium deficiency. Thus far such associations identified have a methodological bias towards being single nucleotide variants, so it is hoped that future genomic projects will be able to determine if other types of mutations play a major role in selenium-related pathologies.

4. Conclusions

The genomics of selenium, selenocysteine and selenoproteins has come a long way since the first prokaryotic and eukaryotic genomes were sequenced two decades ago. Selenoproteomes have been described in a myriad of species, with the homology between selenocysteine and cysteine in proteins being today the preferred method to catalog selenoproteins in new species. Still, selenium is often not exchangeable in proteins and humans and other species have had to adapt to the levels of selenium in their environment, where it ranges from deficient to toxic. These adaptations have relied for the most part on common standing variation in humans but the functional consequences of these variants remain largely unexplored. Current efforts to sequence healthy and diseased individuals promise new insights into the role of both common and rare variants in the regulation and deregulation of the metabolism and homeostasis of selenium. After 200 years of its discovery, genomics is striving to catch up with the knowledge of selenium in chemistry and molecular biology.

Acknowledgements

This study was funded (Sergi Castellano) by the NIHR HS&DR Programme (14/21/45) and supported by the NIHR GOSH BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

References

- Agamy O, Ben Zeev B, Lev D, Marcus B, Fine D, Su D, Narkis G, Ofir R, Hoffmann C, Leshinsky-Silver E, et al. 2010. Mutations disrupting selenocysteine formation cause progressive cerebello-cerebral atrophy. *Am J Hum Genet* 87:538-544.
- Alanne M, Kristiansson K, Auro K, Silander K, Kuulasmaa K, Peltonen L, Salomaa V, Perola M. 2007. Variation in the selenoprotein S gene locus is associated with coronary heart disease and ischemic stroke in two independent Finnish cohorts. *Hum Genet* 122:355-365.
- Anttonen AK, Hilander T, Linnankivi T, Isohanni P, French RL, Liu Y, Simonovic M, Soll D, Somer M, Muth-Pawlak D, et al. 2015. Selenoprotein biosynthesis defect causes progressive encephalopathy with elevated lactate. *Neurology* 85:306-315.
- Ardisson A, Bragato C, Blasevich F, Maccagnano E, Salerno F, Gandioli C, Morandi L, Mora M, Moroni I. 2016. SEPNI-related myopathy in three patients: novel mutations and diagnostic clues. *Eur J Pediatr* 175:1113-1118.
- Bekaert M, Firth AE, Zhang Y, Gladyshev VN, Atkins JF, Baranov PV. 2010. Recode-2: new design, new search tools, and many more genes. *Nucleic Acids Res* 38:D69-74.
- Bera S, Weinberg F, Ekoue DN, Ansenberger-Fricano K, Mao M, Bonini MG, Diamond AM. 2014. Natural allelic variations in glutathione peroxidase-1 affect its subcellular localization and function. *Cancer Res* 74:5118-5126.
- Berry MJ, Banu L, Chen YY, Mandel SJ, Kieffer JD, Harney JW, Larsen PR. 1991. Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3' untranslated region. *Nature* 353:273-276.
- Berry MJ, Martin GW, 3rd, Low SC. 1997. RNA and protein requirements for eukaryotic selenoprotein synthesis. *Biomed Environ Sci* 10:182-189.
- Beutler E. 2006. Hemochromatosis: genetics and pathophysiology. *Annu Rev Med* 57:331-347.
- Canani LH, Capp C, Dora JM, Meyer EL, Wagner MS, Harney JW, Larsen PR, Gross JL, Bianco AC, Maia AL. 2005. The type 2 deiodinase A/G (Thr92Ala) polymorphism is associated with decreased enzyme velocity and increased insulin resistance in patients with type 2 diabetes mellitus. *J Clin Endocrinol Metab* 90:3472-3478.
- Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, et al. 2002. A human genome diversity cell line panel. *Science* 296:261-262.
- Castellano S, Andres AM, Bosch E, Bayes M, Guigo R, Clark AG. 2009. Low exchangeability of selenocysteine, the 21st amino acid, in vertebrate proteins. *Mol Biol Evol* 26:2031-2040.
- Castellano S, Gladyshev VN, Guigo R, Berry MJ. 2008. SelenoDB 1.0 : a database of selenoprotein genes, proteins and SECIS elements. *Nucleic Acids Res* 36:D332-338.
- Castellano S, Lobanov AV, Chapple C, Novoselov SV, Albrecht M, Hua D, Lescure A, Lengauer T, Krol A, Gladyshev VN, et al. 2005. Diversity and functional plasticity of eukaryotic selenoproteins: identification and characterization of the SelJ family. *Proc Natl Acad Sci U S A* 102:16188-16193.
- Castellano S, Morozova N, Morey M, Berry MJ, Serras F, Corominas M, Guigo R. 2001. In silico identification of novel selenoproteins in the *Drosophila melanogaster* genome. *EMBO Rep* 2:697-702.
- Castellano S, Novoselov SV, Kryukov GV, Lescure A, Blanco E, Krol A, Gladyshev VN, Guigo R. 2004. Reconsidering the evolution of eukaryotic selenoproteins: a novel nonmammalian family with scattered phylogenetic distribution. *EMBO Rep* 5:71-77.
- Chambers I, Frampton J, Goldfarb P, Affara N, McBain W, Harrison PR. 1986. The structure of the mouse glutathione peroxidase gene: the selenocysteine in the active site is encoded by the 'termination' codon, TGA. *EMBO J* 5:1221-1227.
- Chapple CE, Guigo R. 2008. Relaxation of selective constraints causes independent selenoprotein extinction in insect genomes. *PLoS One* 3:e2968.

- Cone JE, Del Rio RM, Davis JN, Stadtman TC. 1976. Chemical characterization of the selenoprotein component of clostridial glycine reductase: identification of selenocysteine as the organoselenium moiety. *Proc Natl Acad Sci U S A* 73:2659-2663.
- Cornelis MC, Fornage M, Foy M, Xun P, Gladyshev VN, Morris S, Chasman DI, Hu FB, Rimm EB, Kraft P, et al. 2015. Genome-wide association study of selenium concentrations. *Hum Mol Genet* 24:1469-1477.
- Curran JE, Jowett JB, Elliott KS, Gao Y, Gluschenko K, Wang J, Abel Azim DM, Cai G, Mahaney MC, Comuzzie AG, et al. 2005. Genetic variation in selenoprotein S influences inflammatory response. *Nat Genet* 37:1234-1241.
- Driscoll CA, Macdonald DW, O'Brien SJ. 2009. From wild animals to domestic pets, an evolutionary view of domestication. *Proc Natl Acad Sci U S A* 106 Suppl 1:9971-9978.
- Du XA, Wang HM, Dai XX, Kou Y, Wu RP, Chen Q, Cao JL, Mo XY, Xiong YM. 2015. Role of selenoprotein S (SEPS1) -105G>A polymorphisms and PI3K/Akt signaling pathway in Kashin-Beck disease. *Osteoarthritis Cartilage* 23:210-216.
- Dumitrescu AM, Liao XH, Abdullah MS, Lado-Abeal J, Majed FA, Moeller LC, Boran G, Schomburg L, Weiss RE, Refetoff S. 2005. Mutations in SECISBP2 result in abnormal thyroid hormone metabolism. *Nat Genet* 37:1247-1252.
- Ekoue DN, Bera S, Ansong E, Hart PC, Zaichick S, Domann FE, Bonini MG, Diamond AM. 2017. Allele-specific interaction between glutathione peroxidase 1 and manganese superoxide dismutase affects the levels of Bcl-2, Sirt3 and E-cadherin. *Free Radic Res* 51:582-590.
- Engelken J, Espadas G, Mancuso FM, Bonet N, Scherr AL, Jimenez-Alvarez V, Codina-Sola M, Medina-Stacey D, Spataro N, Stoneking M, et al. 2016. Signatures of Evolutionary Adaptation in Quantitative Trait Loci Influencing Trace Element Homeostasis in Liver. *Mol Biol Evol* 33:738-754.
- Forchhammer K, Bock A. 1991. [Biology and biochemistry of selenium]. *Naturwissenschaften* 78:497-504.
- Forstrom JW, Zakowski JJ, Tappel AL. 1978. Identification of the catalytic site of rat liver glutathione peroxidase as selenocysteine. *Biochemistry* 17:2639-2644.
- Foster CB, Aswath K, Chanock SJ, McKay HF, Peters U. 2006. Polymorphism analysis of six selenoprotein genes: support for a selective sweep at the glutathione peroxidase 1 locus (3p21) in Asian populations. *BMC Genet* 7:56.
- Fu WQ, O'Connor TD, Jun K, Kang HM, Abecasis G, Leal SM, Gabriel S, Rieder MJ, Altshuler D, Shendure J, et al. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants (vol 493, pg 216, 2013). *Nature* 495:270-270.
- Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jorgensen ME, Korneliussen TS, Gerbault P, Skotte L, Linneberg A, et al. 2015. Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science* 349:1343-1347.
- Geller F, Feenstra B, Carstensen L, Pers TH, van Rooij IA, Korberg IB, Choudhry S, Karjalainen JM, Schnack TH, Hollegaard MV, et al. 2014. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. *Nat Genet* 46:957-963.
- Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, et al. 2015. A global reference for human genetic variation. *Nature* 526:68-74.
- Gentschew L, Bishop KS, Han DY, Morgan AR, Fraser AG, Lam WJ, Karunasinghe N, Campbell B, Ferguson LR. 2012. Selenium, selenoprotein genes and Crohn's disease in a case-control population from Auckland, New Zealand. *Nutrients* 4:1247-1259.
- Gladyshev VN, Arner ES, Berry MJ, Brigelius-Flohe R, Bruford EA, Burk RF, Carlson BA, Castellano S, Chavatte L, Conrad M, et al. 2016. Selenoprotein Gene Nomenclature. *J Biol Chem* 291:24036-24040.
- Gladyshev VN, Hatfield DL. 1999. Selenocysteine-containing proteins in mammals. *J Biomed Sci* 6:151-160.
- Gobler CJ, Lobanov AV, Tang YZ, Turanov AA, Zhang Y, Doblin M, Taylor GT, Sanudo-Wilhelmy SA, Grigoriev IV, Gladyshev VN. 2013. The central role of selenium in the biochemistry and ecology of the harmful pelagophyte, *Aureococcus anophagefferens*. *ISME J* 7:1333-1343.

- Gudmundsson J, Sulem P, Gudbjartsson DF, Blondal T, Gylfason A, Agnarsson BA, Benediktsdottir KR, Magnusdottir DN, Orlygsdottir G, Jakobsdottir M, et al. 2009. Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. *Nat Genet* 41:1122-1126.
- Guimaraes MJ, Peterson D, Vicari A, Cocks BG, Copeland NG, Gilbert DJ, Jenkins NA, Ferrick DA, Kastelein RA, Bazan JF, et al. 1996. Identification of a novel selD homolog from eukaryotes, bacteria, and archaea: is there an autoregulatory mechanism in selenocysteine metabolism? *Proc Natl Acad Sci U S A* 93:15086-15091.
- Hamanishi T, Furuta H, Kato H, Doi A, Tamai M, Shimomura H, Sakagashira S, Nishi M, Sasaki H, Sanke T, et al. 2004. Functional variants in the glutathione peroxidase-1 (GPx-1) gene are associated with increased intima-media thickness of carotid arteries and risk of macrovascular diseases in japanese type 2 diabetic patients. *Diabetes* 53:2455-2460.
- Holben DH, Smith AM. 1999. The diverse role of selenium within selenoproteins: A review. *Journal of the American Dietetic Association* 99:836-843.
- International HapMap C. 2005. A haplotype map of the human genome. *Nature* 437:1299-1320.
- Keinan A, Clark AG. 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 336:740-743.
- Kryukov GV, Castellano S, Novoselov SV, Lobanov AV, Zehntab O, Guigo R, Gladyshev VN. 2003. Characterization of mammalian selenoproteomes. *Science* 300:1439-1443.
- Kryukov GV, Gladyshev VN. 2004. The prokaryotic selenoproteome. *EMBO Rep* 5:538-543.
- Kryukov GV, Kryukov VM, Gladyshev VN. 1999. New mammalian selenocysteine-containing proteins identified with an algorithm that searches for selenocysteine insertion sequence elements. *J Biol Chem* 274:33888-33897.
- Lescure A, Gautheret D, Carbon P, Krol A. 1999. Novel selenoproteins identified in silico and in vivo by using a conserved RNA structural motif. *J Biol Chem* 274:38147-38154.
- Liu Z, Reches M, Groisman I, Engelberg-Kulka H. 1998. The nature of the minimal 'selenocysteine insertion sequence' (SECIS) in *Escherichia coli*. *Nucleic Acids Res* 26:896-902.
- Lobanov AV, Fomenko DE, Zhang Y, Sengupta A, Hatfield DL, Gladyshev VN. 2007. Evolutionary dynamics of eukaryotic selenoproteomes: large selenoproteomes may associate with aquatic life and small with terrestrial life. *Genome Biol* 8:R198.
- Lobanov AV, Gromer S, Salinas G, Gladyshev VN. 2006. Selenium metabolism in *Trypanosoma*: characterization of selenoproteomes and identification of a Kinetoplastida-specific selenoprotein. *Nucleic Acids Res* 34:4012-4024.
- Lobanov AV, Hatfield DL, Gladyshev VN. 2009. Eukaryotic selenoproteins and selenoproteomes. *Biochim Biophys Acta* 1790:1424-1428.
- Lobanov AV, Hatfield DL, Gladyshev VN. 2008. Reduced reliance on the trace element selenium during evolution of mammals. *Genome Biol* 9:R62.
- Lopez Herraes D, Bauchet M, Tang K, Theunert C, Pugach I, Li J, Nandineni MR, Gross A, Scholz M, Stoneking M. 2009. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* 4:e7888.
- Mariotti M. 2018. SECISearch3 and Sebastian: In-Silico Tools to Predict SECIS Elements and Selenoproteins. *Methods Mol Biol* 1661:3-16.
- Mariotti M, Guigo R. 2010. Selenoprofiles: profile-based scanning of eukaryotic genome sequences for selenoprotein genes. *Bioinformatics* 26:2656-2663.
- Mariotti M, Lobanov AV, Guigo R, Gladyshev VN. 2013. SECISearch3 and Sebastian: new tools for prediction of SECIS elements and selenoproteins. *Nucleic Acids Res* 41:e149.
- Mariotti M, Ridge PG, Zhang Y, Lobanov AV, Pringle TH, Guigo R, Hatfield DL, Gladyshev VN. 2012. Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLoS One* 7:e33066.

- Mariotti M, Santesmasses D, Capella-Gutierrez S, Mateo A, Arnan C, Johnson R, D'Aniello S, Yim SH, Gladyshev VN, Serras F, et al. 2015. Evolution of selenophosphate synthetases: emergence and relocation of function through independent duplications and recurrent subfunctionalization. *Genome Res* 25:1256-1267.
- Martin-Romero FJ, Kryukov GV, Lobanov AV, Carlson BA, Lee BJ, Gladyshev VN, Hatfield DL. 2001. Selenium metabolism in *Drosophila*: selenoproteins, selenoprotein mRNA expression, fertility, and mortality. *J Biol Chem* 276:29798-29804.
- Meplan C, Hesketh J. 2012. The influence of selenium and selenoprotein gene variants on colorectal cancer risk. *Mutagenesis* 27:177-186.
- Meplan C, Rohrmann S, Steinbrecher A, Schomburg L, Jansen E, Linseisen J, Hesketh J. 2012. Polymorphisms in thioredoxin reductase and selenoprotein K genes and selenium status modulate risk of prostate cancer. *PLoS One* 7:e48709.
- Otero L, Romanelli-Cedrez L, Turanov AA, Gladyshev VN, Miranda-Vizuete A, Salinas G. 2014. Adjustments, extinction, and remains of selenocysteine incorporation machinery in the nematode lineage. *RNA* 20:1023-1034.
- Peng T, Lin J, Xu YZ, Zhang Y. 2016. Comparative genomics reveals new evolutionary and ecological patterns of selenium utilization in bacteria. *ISME J* 10:2048-2059.
- Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, Werner J, Villanea FA, Mountain JL, Misra R, et al. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39:1256-1260.
- Perry JR, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, He C, Chasman DI, Esko T, Thorleifsson G, et al. 2014. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. *Nature* 514:92-97.
- Persson BC, Bock A, Jackle H, Vorbruggen G. 1997. SelD homolog from *Drosophila* lacking selenide-dependent monoselenophosphate synthetase activity. *J Mol Biol* 274:174-180.
- Ratnasinghe D, Tangrea JA, Andersen MR, Barrett MJ, Virtamo J, Taylor PR, Albanes D. 2000. Glutathione peroxidase codon 198 polymorphism variant increases lung cancer risk. *Cancer Res* 60:6381-6383.
- Romagne F, Santesmasses D, White L, Sarangi GK, Mariotti M, Hubler R, Weihmann A, Parra G, Gladyshev VN, Guigo R, et al. 2014. SelenoDB 2.0: annotation of selenoprotein genes in animals and their genetic diversity in humans. *Nucleic Acids Res* 42:D437-443.
- Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, Cotsapas C, Xie X, Byrne EH, McCarroll SA, Gaudet R, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913-918.
- Sadd BM, Barribeau SM, Bloch G, de Graaf DC, Dearden P, Elsik CG, Gadau J, Grimmelikhuijzen CJ, Hasselmann M, Lozier JD, et al. 2015. The genomes of two key bumblebee species with primitive eusocial organization. *Genome Biol* 16:76.
- Santos LR, Duraes C, Mendes A, Prazeres H, Alvelos MI, Moreira CS, Canedo P, Esteves C, Neves C, Carvalho D, et al. 2014. A polymorphism in the promoter region of the selenoprotein S gene (SEPS1) contributes to Hashimoto's thyroiditis susceptibility. *J Clin Endocrinol Metab* 99:E719-723.
- Sarangi GK, Romagné F, Castellano S. 2018. Distinct patterns of selection in selenium-dependent genes between land and aquatic vertebrates. *Mol Biol Evol*, Advanced online publication.
- Schoenmakers E, Agostini M, Mitchell C, Schoenmakers N, Papp L, Rajanayagam O, Padidela R, Ceron-Gutierrez L, Doffinger R, Prevosto C, et al. 2010. Mutations in the selenocysteine insertion sequence-binding protein 2 gene lead to a multisystem selenoprotein deficiency disorder in humans. *J Clin Invest* 120:4220-4235.
- Schoenmakers E, Carlson B, Agostini M, Moran C, Rajanayagam O, Bochukova E, Tobe R, Peat R, Gevers E, Muntoni F, et al. 2016. Mutation in human selenocysteine transfer RNA selectively disrupts selenoprotein synthesis. *J Clin Invest* 126:992-996.
- Schweizer U, Fradejas-Villar N. 2016. Why 21? The significance of selenoproteins for human health revealed by inborn errors of metabolism. *FASEB J* 30:3669-3681.
- Selinus O, Alloway B, Centeno JA, Finkleman R, B., Fuge R, Lindh U, Smedley P. 2005. *Essentials of Medical Geology: Impacts of the Natural Environment on Public Health*. Burlington: Elsevier Academic Press.

- Stadtman TC. 1996. Selenocysteine. *Annu Rev Biochem* 65:83-100.
- Sunde RA, Raines AM. 2011. Selenium regulation of the selenoprotein and nonselenoprotein transcriptomes in rodents. *Adv Nutr* 2:138-150.
- Taskov K, Chapple C, Kryukov GV, Castellano S, Lobanov AV, Korotkov KV, Guigo R, Gladyshev VN. 2005. Nematode selenoproteome: the use of the selenocysteine insertion system to decode one codon in an animal genome? *Nucleic Acids Res* 33:2227-2238.
- Tennesen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, et al. 2012. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* 337:64-69.
- Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, et al. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39:31-40.
- Tugendreich S, Boguski MS, Seldin MS, Hieter P. 1993. Linking yeast genetics to mammalian genomes: identification and mapping of the human homolog of CDC27 via the expressed sequence tag (EST) data base. *Proc Natl Acad Sci U S A* 90:10031-10035.
- Ursini F, Heim S, Kiess M, Maiorino M, Roveri A, Wissing J, Flohe L. 1999. Dual function of the selenoprotein PHGPx during sperm maturation. *Science* 285:1393-1396.
- Walczak R, Westhof E, Carbon P, Krol A. 1996. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA* 2:367-379.
- White L, Romagne F, Muller E, Erlebach E, Weihmann A, Parra G, Andres AM, Castellano S. 2015. Genetic adaptation to levels of dietary selenium in recent human history. *Mol Biol Evol* 32:1507-1518.
- Wilber CG. 1980. Toxicology of selenium: a review. *Clin Toxicol* 17:171-230.
- Wilting R, Schorling S, Persson BC, Bock A. 1997. Selenoprotein synthesis in archaea: identification of an mRNA element of *Methanococcus jannaschii* probably directing selenocysteine insertion. *J Mol Biol* 266:637-641.
- Xiong YM, Zou XZ, Chen Q, Du XL, Liu JR. 2015. Relationship between Gpx1 Pro198leu Polymorphism and Susceptibility of Kashin-Beck Disease. *Value in Health* 18:A638-A638.
- Yao Y, Pei F, Kang P. 2011. Selenium, iodine, and the relation with Kashin-Beck disease. *Nutrition* 27:1095-1100.
- Zhang G, Feenstra B, Bacelis J, Liu X, Muglia LM, Juodakis J, Miller DE, Litterman N, Jiang PP, Russell L, et al. 2017. Genetic Associations with Gestational Duration and Spontaneous Preterm Birth. *N Engl J Med* 377:1156-1167.
- Zhang W, Ramanathan CS, Nadimpalli RG, Bhat AA, Cox AG, Taylor EW. 1999. Selenium-dependent glutathione peroxidase modules encoded by RNA viruses. *Biol Trace Elem Res* 70:97-116.
- Zhang Y, Fomenko DE, Gladyshev VN. 2005. The microbial selenoproteome of the Sargasso Sea. *Genome Biol* 6:R37.
- Zhang Y, Gladyshev VN. 2005. An algorithm for identification of bacterial selenocysteine insertion sequence elements and selenoprotein genes. *Bioinformatics* 21:2580-2589.
- Zhang Y, Gladyshev VN. 2007. High content of proteins containing 21st and 22nd amino acids, selenocysteine and pyrrolysine, in a symbiotic deltaproteobacterium of gutless worm *Olavius algarvensis*. *Nucleic Acids Res* 35:4952-4963.
- Zhang Y, Gladyshev VN. 2008. Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. *PLoS Genet* 4:e1000095.
- Zhong L, Arner ES, Ljung J, Aslund F, Holmgren A. 1998. Rat and calf thioredoxin reductase are homologous to glutathione reductase with a carboxyl-terminal elongation containing a conserved catalytically active penultimate selenocysteine residue. *J Biol Chem* 273:8581-8591.

Zinoni F, Birkmann A, Stadtman TC, Bock A. 1986. Nucleotide sequence and expression of the selenocysteine-containing polypeptide of formate dehydrogenase (formate-hydrogen-lyase-linked) from *Escherichia coli*. *Proc Natl Acad Sci U S A* 83:4650-4654.

Zinoni F, Heider J, Bock A. 1990. Features of the formate dehydrogenase mRNA necessary for decoding of the UGA codon as selenocysteine. *Proc Natl Acad Sci U S A* 87:4660-4664.