The role of mitochondrial-nuclear co-ordination in neurodegenerative disease

Áine Shelley Fairbrother-Browne

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Institute of Neurology

University College London

July 28, 2023

Declaration

I, Áine Shelley Fairbrother-Browne, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Publications arising from this thesis, include:

Fairbrother-Browne, A., Ali, A.T., Reynolds, R.H., Garcia-Ruiz, S., Zhang, D., Chen, Z., Ryten, M., Hodgkinson, A., 2021. Mitochondrial-nuclear cross-talk in the human brain is modulated by cell type and perturbed in neurodegenerative disease. Commun Biol 4, 1262. https://doi.org/10.1038/s42003-021-02792-w

Results and text from these publications have been included in this thesis.

Publications not directly related to this thesis include:

Garcia-Ruiz, S., Zhang, D., Gustavsson, E.K., Rocamora-Perez, G., Grant-Peters, M., **Fairbrother-Browne, A.**, Reynolds, R.H., Brenton, J.W., Gil-Martinez, A.L., Chen, Z., Rio, D.C., Botia, J.A., Guelfi, S., Collado-Torres, L., Ryten, M., 2023. Splicing accuracy varies across human introns, tissues and age. https://doi.org/10.1101/2023.03.29.534370

Chen, Z., Tucci, A., Cipriani, V., Gustavsson, E.K., Ibañez, K., Reynolds, R.H., Zhang, D., Vestito, L., García, A.C., Sethi, S., Brenton, J.W., García-Ruiz, S., **Fairbrother-Browne, A.**, Gil-Martinez, A.-L., Genomics England Research Con-

sortium, Wood, N., Hardy, J.A., Smedley, D., Houlden, H., Botía, J., Ryten, M., 2023. Functional genomics provide key insights to improve the diagnostic yield of hereditary ataxia. Brain awad009. https://doi.org/10.1093/brain/awad009

García-Ruiz, S., Gustavsson, E.K., Zhang, D., Reynolds, R.H., Chen, Z., **Fairbrother-Browne, A.**, Gil-Martínez, A.L., Botia, J.A., Collado-Torres, L., Ryten, M., 2023. IntroVerse: a comprehensive database of introns across human tissues. Nucleic Acids Research 51, D167–D178. https://doi.org/10.1093/nar/gkac1056

Winder, L.A., Parsons, P., Horsburgh, G., Maher, K., Hipperson, H., Wierzbicki, C., Jeffries, A.R., Brown, M.R., **Fairbrother-Browne, A.**, Denise, H., Khalifa, M.S., Bassano, I., van Aerle, R., Williams, R., Farcas, K., Paterson, S., Blackwell, P.G., Burke, T., 2022. Comparison of multiple whole-genome and Spike -only sequencing protocols for estimating variant frequencies via wastewater-based epidemiology (preprint). Public and Global Health. https://doi.org/10.1101/2022.12.22.22283855

Chen, Z., Zhang, D., Reynolds, R.H., Gustavsson, E.K., García-Ruiz, S., D'Sa, K., **Fairbrother-Browne, A.**, Vandrovcova, J., Hardy, J., Houlden, H., Gagliano Taliun, S.A., Botía, J., Ryten, M., 2021. Human-lineage-specific genomic elements are associated with neurodegenerative disease and APOE transcript usage. Nat Commun 12, 2076. https://doi.org/10.1038/s41467-021-22262-5

Abstract

Neurodegenerative diseases (NDs) are progressive neurological conditions that affect millions of people worldwide and yet there are few treatments that can prevent or slow progression. Mitochondrial dysfunction is known to contribute to the pathogenesis of an array of NDs, including Parkinson's Disease (PD) and Alzheimer's disease (AD) for which the causal mechanisms, particularly in the sporadic forms, remain elusive. Mitochondrial processes have been implicated in the pathogenesis of both of these diseases. For mitochondria to carry out their diverse cellular roles, their interaction with the nucleus is essential, owing in part to the nuclear genome encoding $\sim$99% of the mitochondrial proteome. Although this bi-genomic interaction is widely acknowledged, there is little understanding about how this coordination may be important in NDs.

In the first instance, mitochondrial-nuclear coordination in the healthy human brain was assayed by generating gene coexpression profiles across 12 regions, revealing strong regional patterns modulated by cell type and reflecting functional specialisation of the brain. Using an AD case-control paradigm, mitochondrial-nuclear relationships were found to be highly perturbed in cases, particularly through synaptic and lysosomal pathways, implicating energy balance regulation and removal of dysfunction mitochondria in the aetiology or progression of AD.

Mitochondrial dysfunction is thought of as a core component of PD pathogenesis. In light of this, the subsequent aim was to characterise mitochondrial transcriptional profiles in monogenic familial PD, which has known genetic causes and better understood molecular mechanisms. Striking changes in oxidative phosphorylation (OXPHOS) subunit gene expression were observed in PD patients compared to con-

trols, regardless of the underlying genetic cause (monogenic or sporadic). This was particularly marked in monogenic forms. Additionally, unsymptomatic monogenic PD profiles were more similar to those of symptomatic PD cases than to controls. As such, expression of OXPHOS pathway genes was found to be altered in PD, regardless of genetic cause or clinical status. This supports the idea that changes in mitochondrial processes may contribute to disease pathogenesis and points to these changes arising early in the progression trajectory.

Building on these results, the final aim was to study mitochondrial transcriptional control in sporadic PD. Sporadic PD is characterised by its complex genetic origin and cryptic molecular mechanisms. To understand whether nuclear genetic variation modulated mitochondrial transcriptional phenotypes in sporadic PD, mitochondrial-nuclear expression QTLs and post-transcriptional (PT) modification QTLs were generated. This work leveraged the case-control, multi-omic AMP-PD consortium dataset (blood-derived RNA-Seq data from 1483 cases and 965 controls). Nuclear genetic regulation of mtDNA-encoded genes was found to be perturbed in PD, implicating *MAP3K11* and *CD9* in disease pathways that also operate through modulation of mitochondrial processes and pointing to a role for mitochondrial-nuclear co-ordination in the pathogenesis of sporadic PD.

Impact statement

The impact of this PhD thesis is mainly in contributing to our understanding of neurodegenerative diseases (NDs) through the discovery of valuable mechanistic insights. As well as this, a parallel focus has been the development of useful pipelines and tools that can be utilised by the research community to aid wider research efforts.

The thesis resulted in the publication of chapter 2 in *Nature Communications Biology*. As such, some of the work in this thesis has been made available to the wider research community, increasing the impact and reach of the research. The bespoke web tool (MitoNuclearCOEXPlorer) that accompanied this publication enables other researchers to replicate the analyses performed in the paper, download processed data, run their own queries and generate bespoke publication-quality visualisations. As such, this tool has the potential to contribute to research in this field by making the data and methods developed in this work accessible and user-friendly. Furthermore, the pipelines involved in this work are publicly available, promoting scientific integrity through transparency and reproducibility and allowing researchers in the field to verify and build on the analyses performed here. In addition, chapters three and four will shortly be prepared for publication along with the analysis pipelines that were developed for this work.

For the work set out in chapter four, an open-source R package for querying the Ensembl API was developed and publicly released. This has already been utilised by the research community both within academic and industry settings. This package allows researchers to access and utilise Ensembl data in an efficient, user-friendly way that allows integration into R workflows and encour-

ages code and analysis reproducibility. All pipelines mentioned here, along with the MitoNuclearCOEXPlorer web tool R package can be found on my GitHub: github.com/ainefairbrother.

In summary, this thesis has made significant contributions to academia by providing insights into the mechanisms of NDs, as well as developing pipelines and tools that can be utilised by other researchers in the field. The publication of the research in high-impact scientific journals, as well as the public availability of the tools and pipelines developed in this work hope to ultimately contribute to the development of interventions that may cease or prevent neurodegenerative diseases.

UCL Research Paper Declaration Form: referencing the doctoral candidate's own published work(s)

1. **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?**

   Mitochondrial-nuclear cross-talk in the human brain is modulated by cell type and perturbed in neurodegenerative disease

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

   https://www.nature.com/articles/s42003-021-02792-w

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

   Nature Communications Biology

   (d) **Who published the work?**

   Nature Publishing Group

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

   04/11/2021

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

   Aine Fairbrother-Browne, Aminah T. Ali, Regina H. Reynolds, Sonia Garcia-Ruiz, David Zhang, Zhongbo Chen, Mina Ryten, Alan Hodgkinson

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

   Yes

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

   Yes

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

   https://www.biorxiv.org/content/10.1101/2021.02.04.429781v1

   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)**:**

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

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

   (c) **Where is the work intended to be published?**

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

   (e) **Stage of publication:**

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

   A.F.-B., M.R. and A.H. conceived and designed the study. A.F.-B. analysed the data, generated figures, developed the web resource and together with M.R. and A.H. wrote the first draft of the manuscript. A.T.A. processed the GTEx data from raw to read counts. R.H.R. provided code and specificity matrices for the cell-type-specific analysis. S.G.-R. assisted with the technicalities of the web app. D.Z., Z.C., R.H.R., M.R. and A.H. helped guide and troubleshoot analyses. A.F.-B., A.H. and M.R. contributed to the critical analysis of the manuscript.

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

   Chapter 2: "Assaying mitochondrial-nuclear co-ordination using expression correlation in brain tissue"

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: Aine Fairbrother-Browne

Date: 31/03/23

Supervisor/Senior Author signature:

Alan Hodgkinson

Date: 31/03/23

Acknowledgements

First and foremost I would like to express gratitude to those who have donated their brains and bodies to research, without these selfless contributions, this work as well as the progress and advancements made in this field would not have been achievable.

I would like to thank the LIDo programme for granting me a studentship and for giving me the invaluable opportunity to study for a PhD, as well as to the BBSRC for funding. Thanks also to the wider LIDo community and in particular to the cohort of 2018 who have always been there to inject fun and camaraderie.

I could not have undertaken this journey without the wisdom, expertise, patience and kindness of my two supervisors **Dr. Alan Hodgkinson** and **Professor Mina Ryten**. I have been incredibly grateful for the opportunity to be mentored, guided and nurtured by both of you. If I have absorbed even a tiny iota of your brilliance, I will be very lucky indeed. My sincerest thanks to you both and I hope that we will always be in touch.

Equally, I would like to extend my heartfelt thanks to all past and present members of the **Ryten lab** at UCL. You are all brilliant scientists, excellent human beings and wonderful friends who I am fortunate to know. **David Zhang** and **Regina Reynolds**, when I first joined the lab you welcomed me in, took me under your wings and taught me so much (including Tidyverse). **Sonia García-Ruiz**, we have been on this path together since the start — thank you for all the chats, laughs and hugs, I hope we will always be friends. **Emil Gustavsson** thank you for proofreading, for consistently imparting knowledge and experience of which you are an endless fountain, but also for our frequent coffee breaks and food-related conversations, they kept me sane. And of course thanks everyone else for your myriad

support, you're all brilliant humans: **Zhongbo Chen**, **Jon Brenton**, **Kylie Montgomery**, **Ana Luisa Gil Martinez**, **Aaron Wagen**, **Amy Hicks** and also to past and present members of the **Hogkinson lab** at KCL.

Much gratitude and thanks of course to my wonderful and caring LIDo friends **Hanadi**, **Greg**, **Maria** and **Merlin**, with you I have found beautiful life-long friendship.

Words cannot express my gratitude for the love and support of my family. My Mum **Una**, you're an excellent scientist and mother and have always been an anchor to me, my Dad **Paul**, my Grandma **Pat**, my Auntie **Ann-Marie**, and my siblings **Niamh**, **Finn**, **Eibhleann** and **Cass**, you have all been a constant source of fun and encouragement. Thank you for never failing to make me laugh, cheer me up and always being proud of me. My achievements are yours.

Finally, I must thank my endlessly kind, loving, patient, graceful, hilarious, witty partner **Imogen**. You helped me weather all of the storms with boundless empathy, laughter, joy and of course, songs. You are a remarkable, unique and talented individual who I am very lucky to be beside (and thank you for your keen proof reading services!).

Contents

List of Figures

List of Tables

Abbreviations

AD Alzheimer's Disease

AMP-PD Accelerating Medicines Partnership for Parkinson's Disease

BH Bonferroni-Hochberg

CNS Central Nervous System

eQTL Expression Quantitative Trait Locus

ETC Electron Transport Chain

GTEx Genotype-Tissue Expression

GWAS Genome Wide Association Study

meQTL Methylation Quantitative Trait Locus

mRNA Messenger RNA

mtDNAcn Mitochondrial DNA copy number

mtDNA Mitochondrial DNA

nDNA Nuclear DNA

ND Neurodegenerative Disease

NUMTs Nuclear Mitochondrial DNA Segments

OXPHOS Oxidative Phosphorylation

PCA Principal Component Analysis

PC Principal Component

PDBP Parkinson's Disease Biomarkers Program

PD Parkinson's Disease

PMI Post-mortem Interval

PPMI Parkinson's Progression Markers Initiative

PT Post-transcriptional

RIN RNA Integrity Number

RNA-Seq Ribonucleic Acid Sequencing

ROSMAP Religious Orders Study/ Memory and Aging Project

ROS Reactive Oxygen Species

rRNA Ribosomal RNA

SNP Single Nucleotide Polymorphism

snRNA-Seq Single Nuclear RNA Sequencing

TPM Transcrips Per Million

tRNA Transfer RNA

WGS Whole Genome Sequencing

Chapter 1

Introduction

Neurodegenerative diseases (NDs) are devastating progressive neurological disorders that are associated with debilitating physical and mental symptoms. Collectively, they affect millions of people worldwide and yet we lack treatments that would prevent or slow their progression. To do this, we will need to first understand the genetic and molecular drivers of disease risk, onset and progression. Mitochondrial dysfunction has been implicated in the aetiology of diseases across the ND spectrum, and while efforts have been made to understand the precise mechanisms involved, a great deal remain unclear. Core to the the varied and essential roles carried out by mitochondria is a dependence on the nuclear genome which provides $\sim$99% of the mitochondrial proteome. This points to an essential functional role for the integrity of mitochondrial-nuclear coordination. As such, this thesis aims to further understand the role of mitochondria in ND through the lens of this coordination and to explore whether pathological consequences result from the alteration of this relationship.

1.1 Mitochondrial biology and the mitochondrial-nuclear relationship

1.1.1 A brief history and description of the mitochondrion

The mitochondrion was at one time a free-living organism until it encountered and was engulfed by an obligate aerobic or perhaps prokaryotic host, such are the

two main theories, in a primordial endosymbiosis event that took place approximately ∼1.5-2 billion years ago[1]. The eukaryotic cell of the present day is the chimeric outcome, comprising numerous mitochondria, each of which contain multiple copies of their own genome. Over evolutionary time, mitochondrial DNA (mtDNA) encoded genes have been transferred to the nuclear genome, resulting in a much diminished and nuclear-dependent mitochondrial genome[2]. Owing to this, the double-stranded circular human mitochondrial genome of the present is comprised of just 16,569 DNA base pairs, containing 37 genes, 13 of which encode proteins, 2 encode ribosomal RNAs (rRNAs) and 22 encode transfer RNAs (tRNAs). The 13 protein products are components of the electron transport chain (ETC) which are embedded in the inner mitochondrial membrane (IMM) and together with nuclear DNA (nDNA) encoded proteins, carry out oxidative phosphorylation (OXPHOS). In this process, adenosine triphosphate (ATP), the energy storage molecule of the cell, is generated from glucose. As such, within the eukaryotic cell reside two separate but intrinsically linked genomes.

1.1.2 The human mitochondrial transcriptome and epitranscriptome

Reflective of its prokaryotic origins, the human mitochondrial genome is transcribed in a strand-specific, polycistronic manner from both the heavy and light strands[3,4]. 12 protein coding genes, 14 tRNAs, both rRNAs and a variety of non-coding RNA species are encoded on and transcribed from the heavy strand, whereas just one protein coding gene (*MT-ND6*) and one ncRNA (non-coding RNA) are transcribed from the light strand. Although recent evidence challenges this canonical model, highlighting the presence of small open reading frames in the mtDNA coding for microproteins[5]. Furthermore, the variety of mtDNA-encoded non-coding RNA species is an evolving area of study and to date long non-coding, circular, double stranded and small RNAs have been identified[6]. The displacement loop (D-loop) in the so-called control region of the mtDNA and is a ∼1100 nucleotide non-coding span, containing regulatory elements responsible for replication and transcription, including the light strand promoter (LSP) and heavy strand promoter (HSP). In

humans, three main nuclear DNA (nDNA) -encoded proteins are essential for mitochondrial transcription: (i) mitochondrial transcription factor A, TFAM, binds upstream of the transcription start site (TSS), (ii) TFAM recruits the mitochondrial RNA polymerase, POLRMT, which binds to the promoters and finally, (iii) the mitochondrial transcription elongation factor, TEFM, increases transcriptional processivity[7,8,9].

The mitochondrial 'epitranscriptome' describes the array of mitochondrial post-transcriptional (PT) RNA modifications which act as chemical and topological augmentations of nucleotide bases (A, C, G and U). They are known to be highly conserved across species, prevalent and varied[10,11,12,13]. These modifications enable dynamic fine-tuning of RNA function and as such have a central role in transcription[12]. However, the exact function of many of these remains elusive[11].

In mitochondria, a large proportion of PT RNA modifications occur on mitochondrial tRNAs (mt-tRNAs) for which 18 types at 137 positions have been catalogued to date[14]. During mitochondrial translation, 22 mtDNA-encoded tRNAs are responsible for decoding 60 codons[14]. This is in stark contrast to the >500 tRNAs encoded by the nuclear genome used to decode 61 codons. 22 is the biological minimal set required and is permitted by the existence of tRNAs hosting differential modification at the so-called 'wobble position' of the anticodon[14]. At this position, a lack of modification confers promiscuity of recognition to the tRNA, permitting it to read any base at codon position three, conversely, presence of a modification at the wobble position prevents this[14]. As such, specific PT modifications confer expansion of the decoding capacity of the minimal tRNA set.

The downstream impact of each modification is dependent on its type and location and as such, modifications at other positions are known to impact other aspects of tRNA function. The product of mitochondrial transcription, the polycistron, is punctuated by tRNAs which are folded into secondary cloverleaf-shaped structures[15]. Methylation (m1A/G) at the 9^{th} position of the punctuating tRNAs — termed 'p9' sites — is known to be critical to the stability of the tRNA secondary structure. This is important for PT processing functionality, such as excision of

transcripts from the polycistron[16,17]. In this way, PT modifications are instrumental in the observed deviation of mitochondrial actual transcript levels from the 1:1 ratio expected to arise from polycistronic transcription[3].

Given the importance of PT RNA modifications to mitochondrial functionality, it is unsurprising that their dysfunction has been implicated in a number of diseases. For example, in MERRF (myoclonic epilepsy with ragged red fibres) patients, it was found that in individuals carrying the mtDNA m.8344A>G point mutation concurrently lacked methylation at position 58 in *MT-TK*[18]. The authors were able to recover the damaged translational phenotype through overexpression of the methyltransferase TRMT61B, showing that the primary driver of the pathology was indeed elimination of modification in *MT-TK*[18]. Additionally, Across 12 types of cancer, significant changes to m1A/G methylation levels in mt-tRNAs were observed between normal and tumour samples[19]. More recently, a study analysing m1A/G methylation at p9 sites in individuals diagnosed with Alzheimer's disease (AD), progressive supranuclear palsy (PSP) and pathological ageing found similar methylation profiles between AD and PSP cases, pointing to a shared pathological mitochondrial methylation profile[20].

Innovative methodology permitting the inference of mitochondrial PT methylation from bulk RNA sequencing (RNA-Seq) data has enabled high-throughput efforts seeking to characterise their variation across large numbers of healthy individuals and across multiple tissues[21]. This method relies on the principle that mismatched bases in RNA-Seq data arising from mtDNA indicates the presence of a methylation PT modification at that site. During library preparation, a reverse transcriptase enzyme converts RNA to cDNA. Upon encountering a PT modification, however, the enzyme can stall, fall off entirely, or incorporate a random non-complementary nucleotide base at that position[22]. It has been shown that this signature of mismatched bases from the reference is a good proxy for methylation rate, displaying replicability across independent RNA-Seq experiments[21,19,22].

All considered, tools are available that permit high-throughput profiling of mitochondrial PT modifications. This will be beneficial for furthering our understand-

ing disease aetiology in general, but may be particularly relevant for diseases where mitochondria are known to have a central role.

1.1.3 Mitochondrial copy number and heteroplasmy

An important feature of mitochondria is their quantity. Each cell contains many mitochondria and each mitochondrion contains multiple copies of their own genome. The number of copies of the mitochondrial genome per nucleated cell is known as the mitochondrial copy number (mtDNAcn). The mtDNAcn has been found to vary vastly not only across cell types, but temporally within a cell type[23]. Cell types with high energy requirements such as neurons and cardiomyocytes can have mtDNAcn many orders of magnitude higher than less energy dependent cell types such as hepatocytes, for example. Increases and decreases in mtDNAcn have both been linked to pathology, where the former has been posited as a mechanism acting to compensate for low cellular respiratory capacity brought about by prevailing mitochondrial dysfunction[24,25,26]. Indeed, a recent study implicated mtDNAcn in ND, finding that increased mtDNAcn in blood is associated with decreased ND risk[27]. That this has been observed in blood, and that mtDNAcn is a readily measureable phenotype highlights the potential utility for mtDNAcn as a marker of disease, particularly for NDs where progression begins many years prior to symptom onset. Importantly, however, it has recently been suggested that a number of studies finding associations between common diseases and blood mtDNAcn are confounded by blood cell type composition[28]. Wherein the findings of four such studies could be replicated but were subsequently lost when correction for blood cell type composition was applied[28]. As such, much remains to be understood about the significance of mtDNAcn with respect to disease states.

As there are multiple copies of the mtDNA in each cell, mtDNA variants can exist on a spectrum across mtDNA molecules. A given variant can occur in all copies (homoplasmy) or in a proportion of copies (heteroplasmy). Heteroplasmy levels can differ from cell to cell within a tissue, from tissue to tissue within an organ, from organ to organ and from person to person[29]. Pathogenic variants in the mtDNA have been identified in the protein, tRNA and rRNA coding genes.

Severe variants are not tolerated in the homoplasmic state, but rather can exist in the heteroplasmic state, causing OXPHOS dysfunction when a tolerance threshold is exceeded in the cell[23,30]. This threshold is variant and cell type specific, but for certain mtDNA deletions this has been found to be ∼50-60%[23,31].

Severe maternally inherited pathogenic variants in mtDNA are known to cause respiratory chain dysfunction and the associated conditions are referred to as primary mitochondrial diseases, wherein the mtDNA variant is essential for the disease to manifest. These diseases are heterogeneous both in their clinical presentation and genetic underpinnings[29]. They can be caused by different combinations of plasmies and variant types, ranging from small effect homoplasmic mtDNA point mutations to large effect heteroplasmic mtDNA deletions, where the level of heteroplasmy has been found to correlate with the severity of the clinical phenotype[29]. In contrast to these inherited variants of large effect, small effect somatic mtDNA variants have been implicated in common complex diseases. As the somatic mutational burden increases, this has been shown to result in mitochondrial dysfunction[32]. At the level of the tissue, if many cells are mitochondrially non-functional, this can contribute to pathology. In Parkinson's disease (PD) for example, high levels of mtDNA deletions have been identified as a component of impaired cellular respiration and selective neuronal loss in the substantia nigra[33,34].

1.1.4 The diverse roles of mitochondria

The canonical role of mitochondria is as the generator of cellular energy via the metabolism of glucose. However, it is now known that the mitochondria fulfil other vital and interlinked roles. Several studies have suggested that the mitochondria have a role in local and cytoplasmic calcium ion (Ca^{2+}) homeostasis. Recent thought posits that this is heavily dependent on the cell type, the presence of other Ca^{2+} removal mechanisms and mitochondrial positioning[35]. The importance of the latter was found to be particularly marked in neurons, underscoring not only the cell type specificity of the mitochondrial role in Ca^{2+} homeostasis but also pointing to the importance of cell type specific mitochondrial function[35,36]. Linked to the mitochondrial role in Ca^{2+} homeostasis is their role in apoptosis, also known as

the cell death pathway. Triggered by a number of effectors, of which one is Ca^{2+}, mitochondrial permeability increases, releasing cytochrome c into the cytosol. This in turn activates caspases which cause mass proteolysis, leading to cell death[35,37]. Further to this, mitochondria are also known to have a role in triggering cascades involved in the innate immune response, of which apoptosis is one component. This is thought to be achieved through mitochondrial stress signalling in response to environmental insults such as toxicity or pathogen infection[38]. The cellular roles for mitochondria go beyond those mentioned here, however, it is clear that mitochondria have complex and diverse functionality in the cell and it is likely that a number of these remain elusive.

1.1.5 Mitochondrial specialisation

Mitochondrial morphology, function and distribution are heterogeneous across cell types, and even within sub-compartments of the same cell[39,40,41]. Mitochondria form interconnected networks within the cell, which are regulated and transported by microtubules and are capable of undergoing fission and fusion events, leading to the formation of diverse structures such as tubular, rod-like, and spherical forms[42]. A remarkable study focusing on brain cell types utilised MITO-Tag mice expressing an outer mitochondrial membrane (OMM) -targeted green fluorescent protein permitting purification of whole mitochondria from specific cell types[41,39]. Through this approach, Misgeld and colleagues compared the mitochondrial proteomes of inhibitory neurons, excitatory neurons and glia extracted from an adult mouse brain. $\sim$15% of the mitochondrial proteome was found to be differentially regulated between these cell types. Interestingly, they observed an enrichment of peroxisomal proteins in mitochondria extracted from astrocytes, reflecting astrocytic fatty acid consumption[41]. As such, the authors demonstrate cell type specific functionality and adaptations amongst these mitochondria[41]. Intra- and inter-cellular heterogeneity of mitochondria has significant implications for diseases of the brain that involve mitochondrial dysfunction, particularly for conditions where selective vulnerability of neuronal populations is a contributing factor such as certain NDs.

1.1.6 Mitochondrial-nuclear coordination

Given the mitochondrion's evolutionary past, its varied and complex cellular roles and dependence on the nucleus, it stands to reason that mechanisms for mitochondrial-nuclear coordination have arisen at multiple levels of organisation. The central — and most well understood — purpose of mitochondrial-nuclear coordination is to quickly and flexibly manage cellular bioenergetic demand[43]. This coordination operates in both directions.

One element of this coordination is mitochondrial retrograde signalling, providing information to the nucleus about their bioenergetic status, primarily encoded through molecular and biochemical signals. For example, reduced mitochondrial electron transport decreases mitochondrial membrane potential, resulting in the release of cytochrome c into the intermembrane space and then into the cytoplasm, triggering a nuclear apoptotic response[43,44]. Another example of this is when the cellular AMP:ATP ratio increases — indicating slower ATP generation — this is known to activate PGC-1α, the nuclear-encoded stimulator of mitochondrial biogenesis[43]. These examples demonstrate the nuclear response to mitochondrial retrograde signalling, involving the activation of nuclear processes, namely mitochondrial quality control which encompasses numerous complex cascades such as biogenesis, mtDNA repair and mitophagy.

Although this is the canonical view, mitochondrial-nuclear coordination does not exist simply as an anterograde-retrograde signalling loop. Recent work has pointed to the involvement of double localisation of mitochondrial-nuclear transcription factors[45,7]. Indeed, regulators of nuclear gene transcription have been found to bind mtDNA, pointing to the involvement of general cellular expression regulators in mtDNA gene expression and raising the possibility that a portion of these are involved in coordination of mitochondrial-nuclear expression[45]. In addition to colocalised regulatory factors, another level at which coordination operates — though notably not through physical colocalisation — is highlighted by the existence of the nuclear-encoded master-regulator of mitochondrial function, PGC-1α, which regulates transcription of nDNA-encoded OXPHOS components[45].

Around this core structure of coordination exist further nuanced and complex mitochondrial-nuclear interactions. Indeed, recent work has uncovered nuclear genetic variants significantly associated with changes in mitochondrial expression[46], mitochondrial PT modification rate[46], mtDNA abundance[28] and mtDNA heteroplasmy[28].

Within this, it is important to acknowledge the challenges that exist in understanding the mitochondrial-nuclear relationship, particularly with respect to assessing the mitochondrial transcriptome. As discussed in section 1.1.3, mitochondria are heterogeneous organelles. The mitochondrial genome can vary between human populations (through mitochondrial haplogroups), between individuals (through mtDNA variants) and between tissues within an individual (mtDNA homoplasmy, mtDNA heteroplasmy and mtDNAcn). These sources of heterogeneity have the potential to confound the interpretation of mtDNA-derived expression data if not appropriately controlled for. For example, cell type is a large driver of variation with respect to mtDNAcn, and as such, in bulk RNA-Seq data, correcting for inter-individual differences in cell type proportions has been shown to control for mtDNAcn by proxy[28]. To control for ancestry-driven mitochondrial genetic differences, one may consider utilising a cohort with homogeneous ancestry, or if this is not possible, including population substructure covariates in the downstream model.

Taken together, much remains to be discovered with respect mitochondrial-nuclear coordination, how it operates across different tissues and cell types, and its relevance to NDs. Future studies should aim to address these gaps — along with carefully considering potential confounders of the data — to improve our understanding of this fundamental process and its potential implications in human health and disease.

1.2 Neurodegenerative disease

According to global population figures published by the the UN, in 2019 1 in 11 people were aged 65 or over, however, by 2050, this is predicted to be 1 in 6. As the population ages, the prevalence of adult neurodegenerative disease (ND) such

as AD and PD is rising[47]. This thesis will consider AD and PD, the two most prevalent NDs.

1.2.1 Alzheimer's disease

AD is the most common cause of dementia, affecting over 6 million people in the United States alone (Alzheimer's Association, 2023). Two thirds of individuals diagnosed with AD are female and the majority of AD occurs in individuals over the age of 65. The latter is referred to as late-onset AD (LOAD) and constitutes $\sim$98% of cases, while the remaining $\sim$1-2% of cases are considered early-onset AD (EOAD)[48]. EOAD is associated with autosomal dominant inheritance of a large effect variant in a single gene and is characterised by rapid disease progression, while LOAD is generally sporadic or idiopathic and associated with protracted progression. In typical AD cases learning and memory deficits occur early on, followed by impaired attention, executive function, speech and language, movement, motor planning, recognition and behavioural changes[48].

Until recently, disease modifying treatments for AD remained elusive, however, a next-generation anti- amyloid-beta (Aβ) antibody, lecanemab, recently showed appreciable efficacy with respect to clinical outcomes[49,50]. In the study participants, cognitive decline in those given lecanemab was found to decrease by 27% compared to a placebo[49]. In January of 2023, this drug was approved for use in the United States by the FDA. However, concerns remain about whether the observed drug-placebo difference of 0.45 on the 18-point Clinical Dementia Rating-Sum of Boxes (CDR-SB) scale is clinically meaningful[49]. Previous work has estimated minimal clinical importance to be a change of 0.98 to 1.63, dependent on the level of cognitive impairment[49]. Additionally, it has been suggested that this drug and related anti-Aβ antibody -based drugs may cause accelerated brain volume loss[51]. However, this trial represents a potentially promising development and ongoing trials will reveal the ability of this drug to affect disease onset in addition to progression[49]. Given this, a deeper understanding of AD aetiology will be important for elucidating further targets that may permit improvement and development of disease modifying drugs. In parallel, this will facilitate the identifi-

cation of individuals that may benefit from such interventions and the improvement of diagnostic accuracy[52] through biomarker discovery.

1.2.1.1 The genetic architecture of Alzheimer's disease

Genetic architecture is a term that encompasses the characteristics of the genetic variation which confer heritability to a trait, where heritability defines the proportion of variance in the trait attributable to genetic variation. Broad-sense heritability estimates for AD are reported to be 59-78%, whereas narrow-sense heritability estimates are study-variable, but reported to be 3-31%[52,53]. A recent review by Andrews and colleagues included an elegant visualisation of AD genetic architecture, which is also included here (see Figure 1.1), conceptualising it as a two-dimensional space defined by minor allele frequency and effect size. At the extreme left of the allele frequency axis sit the rare monogenic gene variants with high penetrance, namely those found in *PSEN1*, *PSEN2* and *APP*. The remainder of AD causal variants are common, with higher minor allele frequencies and low effect sizes, but cumulatively contribute substantially to disease susceptibility or risk[54].

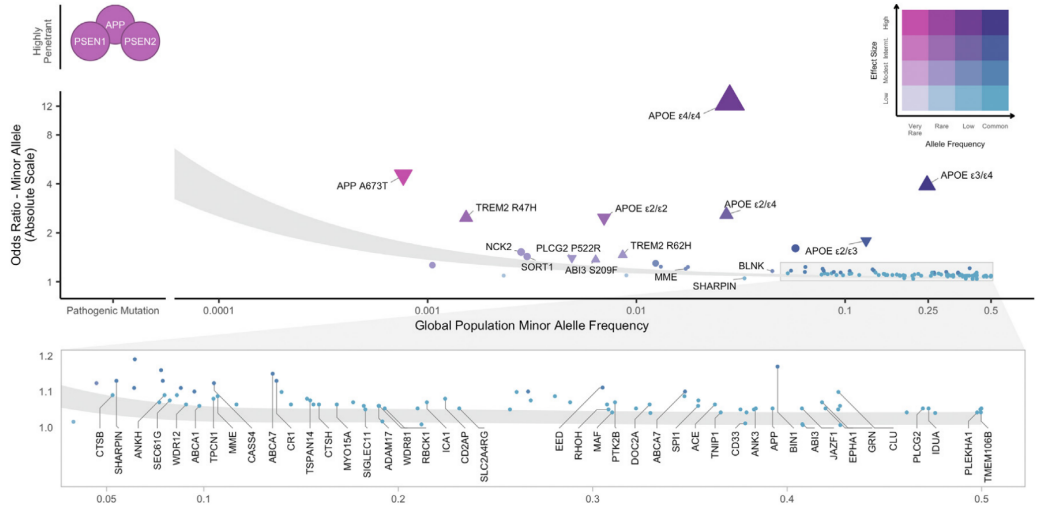

Figure 1.1: The genetic architecture of AD taken from a recent review by Andrews et al., 2023. This figure reports odds ratios on the absolute scale, where the triangles indicate directionality for *APOE* genotypes and labelled loci indicate candidate causal genes from Bellenguez et al., 2022

1.2.1.2 Alzheimer's disease pathways and mechanisms

The discovery of the amyloid precursor protein (*APP*), presenilin 1 (*PSEN1*), and presenilin 2 (*PSEN2*) variants as causal for AD combined with the identification of the epsilon-4 allele of the *APOE4* gene as a risk factor allowed the first major insights into AD pathology[55]. Variants in *APP* prevent its proper proteolysis by secretases, giving rise to populations of Aβ peptide lengths that are more prone to aggregation. These longer peptides are found in Aβ plaques in the brains of individuals with AD. Similarly, variants in *PSEN1* and *PSEN2*, components of the secretase complex, affect Aβ cleavage leading to accumulation-prone Aβ peptide length ratios[55]. The epsilon-4 allele of *APOE* reduces the efficiency of *APOE*-Aβ binding and Aβ aggregation clearance. As such, discovering and understanding these variants and affected processes implicated Aβ and the amyloidogenic pathway in AD pathogenesis, leading to the key theory positing neuronal amyloid plaque formation as central to neurodegeneration in AD.

Increasingly powerful GWA studies have been permitted by improvements in sequencing technology and the availability of increasingly large case/control cohorts, discovering many more loci associated with sporadic AD. Two recent AD GWA studies, released just six months apart, together identified 90 independent variants, implicating 75 loci in AD, 42 of which were novel[54,56,57]. Efforts to prioritise candidate genes arising from these studies found that $\sim$50% of AD loci contained genes involved in myeloid function[54]. Other processes implicated through GWAS findings include lipid transport, inflammation, innate immunity and endocytosis[55]. Interestingly, multiple lines of evidence point to a role for mitochondria as mediators, drivers or as contributors to AD pathogenesis[58]. Current hypotheses posit mitochondrial dysfunction as a consequence of AD pathology, primary agents whose bioenergetic failure sets the stage for downstream pathologies, or as secondary enhancers of primary amyloid pathology[58]. In light of this, further understanding mitochondrial biology and the alterations of mitochondria in disease states may help shed light on its role in AD pathology.

1.2.2 Parkinson's disease

PD is the second most prevalent ND, affecting ∼1 million people in the US alone (Parkinson's Foundation, 2023). Rare monogenic forms of PD (∼5-10% of cases) are associated with an earlier age of onset, whereas the common sporadic form (∼90-95% of cases) is associated with an age of onset approximately of over 60 (Parkinson's Foundation, 2023). Interestingly, PD is twice as common in men, and its incidence appears to vary across populations with respect to genetic ancestry and environment, although it is unclear whether unidentified factors are confounding this[59]. Additionally, environmental factors, independent of genetic predisposition, are thought to modulate PD risk, such as, smoking, alcohol intake, diet and pesticide exposure[59].

First identified by James Parkinson in 1817, the most characteristic diagnostic features of PD are motor symptoms, namely tremor, rigidity and akinesia[60]. The non-motor symptoms can also be very severe, however, and can present during the long prodromal phase. These include cognitive impairment, sleep disorder, depression, anosmia and dysfunction of the autonomic nervous system[59]. PD is pathologically characterised by a loss of dopaminergic neurons in the substantia nigra (SN) pars compacta commonly accompanied by the presence of α-synuclein containing protein aggregates known as Lewy bodies in the midbrain[61]. The dopamine agonist, levodopa, remains the central treatment for PD[62]. Levodopa is converted to dopamine in the peripheral and central nervous systems, activating dopamine receptors, and acting to control motor symptoms[62]. However, it is fraught with issues such as side-effects and low bioavailability and notably does nothing to modify disease progression[62]. Indeed, no disease modifying drugs for PD have been identified at present.

1.2.2.1 The genetic architecture of Parkinson's disease

PD is a genetically heterogeneous disease whose genetic landscape can be thought of as a continuum, similar to that set out for AD in figure 1.1. Rare, highly penetrant, variants of large effect exist at one end of the continuum and at the other are common variants of small effect which contribute to disease risk[63].

Family pedigrees allowed the elucidation of rare PD causing variants, the first of which was the G209A missense variant in *SNCA* identified in 1997[64]. Since then, rare variants in 20 genes have been found to be causative for PD, though not all have been verified through functional studies[61]. Although the rare monogenic familial forms are traditionally characterised by a single causal variant, it is likely that on an individual level, the disease is influenced by other genetic and non-genetic factors[61]. This is evidenced by the existence of highly penetrant variants which do not consistently manifest in the disease and within disease pathogenesis where patients display heterogeneity of symptom onset, progression and symptom severity, pointing to the contribution of factors beyond the single inherited variant[61].

Individuals with sporadic PD are those who do not possess a monogenic familial variant. Since the advent of the GWA study paradigm in 2005, a multitude of PD GWA studies have yielded increasing numbers of genetic risk loci. Indeed, the most recent GWAS meta-analysis detected 90 independent risk signals, of which 38 were novel[61,65]. The portion of disease risk attributable to genetics — the heritability — of PD is estimated to be $\sim$22%, and it is thought that a portion of this can be explained by currently identified GWAS loci[65]. As such, a portion of the heritability of PD remains to be elucidated, and to this end, future GWA studies will likely leverage larger sample sizes and efforts to interpret the biology underlying GWAS associations will be vital.

It is, however, important to acknowledge the distinct challenges that still remain in this arena, the largest of which is biological interpretation, which is compounded by a number of factors[66]. Firstly, GWAS signals often identify genomic regions that are non-coding or regulatory. As such, making biological interpretations of these is complex and requires integration of additional data types such as RNA-Seq and chromatin contact (Hi-C) data which aid in the elucidation of functional mechanisms. Secondly, it has been shown that GWAS-identified variants individually explain very small proportions of total trait variation, are associated with multiple traits and are often proximal and in linkage disequilibrium (LD) with a large number of other variants[66]. As such, determining the causal variant and

identifying the candidate gene or mechanism can be highly challenging. Thirdly, GWAS-detected associations may be heterogeneous across human populations with different ancestries which may be a confounding factor, particularly in GWA studies that leverage large participant cohorts or meta-analyses. Future studies will aim to address these challenges through mutli-omic data integration and improvement of statistical frameworks for fine-mapping. In addition, non-European ancestral groups must be studied as opposed or removed from cohorts, or their heterogeneity corrected for, as they may reveal novel biology. To this end, new efforts are being made by consortia such as The Global Parkinson's Genetics Program (GP2) which aims to leverage diverse human populations to further understand the genetic architecture of PD.

1.2.2.2 Parkinson's disease pathways and mechanisms

To date, a number of key pathways in PD have been implicated through monogenic familial PD genes. For example, *DJ-1*, *PRKN* and *PINK1* are known to have roles in mitochondrial pathways, namely mitophagy. Likewise, lysosomal function has been implicated through *GBA*, *LRRK2* and *VPS35*. Despite this, work is still underway to characterise the complex molecular mechanisms that connect pathogenic genetic variation and clinical manifestations. A key application of this is to understand whether any of the mechanisms discovered here have relevance to sporadic PD, whose aetiology remains comparatively cryptic. The detection of familial PD - associated loci in sporadic PD GWA studies indicates overlap in gene perturbation between the forms, indicating potentially common pathways to disease[65].

A major interest and focus in PD is mitochondrial dysfunction. The link between PD and mitochondria was first established when Langston and colleagues observed that MPTP exposure resulted in chronic PD due in part to its role in mitochondrial complex I activity inhibition[67]. Since this discovery, many more links have emerged, firmly implicating mitochondrial dysfunction in PD pathogenesis. Defective mitochondrial respiration has been observed in PD disease states, mtDNA variants have been found to cause OXPHOS defects in the substantia nigra and α-synuclein accumulation in mitochondria has been found to impair complex I

function[68,69,34]. Furthermore MitoPark murine models which have depleted mitochondria transcription factor A develop defective OXPHOS of the substantia nigra dopaminergic neurons[70]. These and many other studies point to mitochondrial dysfunction as central to PD pathogenesis, however the mechanisms underlying this are far from clear and much work remains to elucidate these.

1.3 Thesis context and aims

It is clear that ND is a major societal problem within which mitochondrial dysfunction is heavily implicated. Further understanding of a key dynamic of mitochondrial function — mitochondrial-nuclear coordination — may be important to gain a full understanding of mitochondrial dysfunction in this context. This thesis aims to achieve this through through leveraging transcriptomic data, taking advantage of the large RNA-seq datasets emerging in the field of ND. Mitochondrial-nuclear cross-talk has been studied previously, but in a limited array of tissues, often in healthy control individuals only, utilising small sample numbers and not in a ND context. As such, the broad aims of this thesis are 3-fold:

(i) To characterise mitochondrial-nuclear cross-talk across the healthy human CNS;

(ii) To study mitochondrial-nuclear cross-talk perturbation in AD and PD;

(iii) To use monogenic PD participant data to gain insights into the role of mitochondrial transcription in PD pathogenesis, across different forms of PD and at different stages of disease.

Chapter 2

Assaying mitochondrial-nuclear co-ordination using expression correlation in brain tissue

2.1 Introduction

The central nervous system (CNS) and its composite cell types possess a distinct metabolic and physical architecture, making its mitochondrial-nuclear relationship of potential importance to further understanding CNS pathologies. Tissues of the CNS are not only highly energetically demanding, consuming 20% of the body's total energy supply[71], but heterogeneous in their requirements, with wide variation in energy demands across their constituent cell types[72,73]. To meet this energy demand, cells of the CNS are highly dependent on mitochondria. However, mitochondria cannot fulfil this role independently of the nucleus. Although mitochondria have their own compact genomes encoding sub-units of the electron transport chain (ETC) (13 genes), $\sim$99% of the proteins required for normal mitochondrial function are encoded in the nucleus (1136 nuclear genes as per MitoCarta3.0, an inventory of mitochondrial proteins)[74,75]. Translation of these nuclear components is carried out by cytoplasmic ribosomes, and protein products are imported into mitochondria. Owing to their association with the nucleus, the mitochondria are fully resourced for their role in energy production, as well as for their other key functions

such as calcium buffering, cellular signalling and apoptosis[76]. Thus, we can see that the efficacy of this bi-genomic system depends on continual coordination of transcriptional activity across the mitochondrial and nuclear genomes.

Certain features of neurons make this coordination challenging, with three that stand out as being most relevant in this context. The first is that neurons are simultaneously highly OXPHOS (oxidative phosphorylation) -dependent, and particularly vulnerable to oxidative stress due to their low anti-oxidant capabilities[77,78]. Intrinsically coupled to the OXPHOS rate is the rate of reactive oxygen species (ROS) production[79]. This is caused by electron leakage from the ETC and manifests as an abundance of toxic hydrogen peroxide and hydroxyl radicals[80]. To manage levels of ROS, the mitochondria utilise a 'ROS defense system' (RDS), composed of ROS-extinguishing enzymes such as isocitrate dehydrogenase, malic enzyme, and transhydrogenase[81]. Failure of this system is thought to damage the mitochondrial DNA (mtDNA), and in turn contribute to accelerated ageing, and by extension, increased risk of neurodegenerative disease (ND)[82].

The second feature is that neurons are terminally differentiated cells, and so their integrity must be maintained for an entire lifespan. As a result, they are highly reliant on autophagic processes for removal of dysfunctional organelles as well as misfolded and aggregated proteins[83]. Maintenance of mitochondrial fidelity requires the autophagic process of mitophagy. The importance of mitophagy is most clearly demonstrated by work showing that key genes in the pathway, including *PINK1* and *PARK2* which encode proteins that tag damaged mitochondria for mitophagic removal, when mutated cause familial Parkinson's disease (PD)[84,85]. Again, this demonstrates a high dependence on nuclear transcription and import for an essential mitochondrial process, and one that is inseparable from neuronal maintenance.

The third important feature of neurons is that they have a unique and highly specialised architecture. Neurons require mitochondrial distribution to satisfy heterogeneous local energy demands used for restoration of ion gradients and axonal transport, as well as for other important mitochondrial roles such as calcium buffer-

ing, neurotransmitter metabolism and ROS generation[86,87]. As well as mitochondrial mobility, it is also necessary for the cell to restrict their movement. Electron microscopy experiments have been able to observe that mitochondria are tethered to sites of high energy requirement, such as vesicle release sites at synapses[88]. Linked to this, there is evidence to suggest that localised protein synthesis occurs in neurons, requiring transport of nuclear DNA (nDNA) -encoded mitochondrial gene transcripts to distal mitochondrial locales[89]. We therefore see a requirement for a dynamic mitochondrial distribution system, where it is necessary for mitochondria to be synthesised and transported, but also a requirement for local coordination of cellular components and mitochondria. This points to an equally dynamic and complex nuclear-mitochondrial coordination system as key to upholding this.

Given the intricacy and scale of mitochondrial-nuclear coordination required in human brain tissue, there is ample opportunity for dysfunction. In neurons, failure of coordinated mitochondrial clearance and biosynthesis contributes to disease pathogenesis. This can be seen in the aetiology of PD, where variants in *PINK1* and *PARK2* are associated with autosomal-recessive PD and their protein products have been implicated not only in mitophagy, but also mitochondrial biogenesis[90,91]. However, pathology of the mitochondrial biogenesis and quality control pathways is not unique to PD. Analyses of brain samples from individuals with Alzheimer's disease (AD) have shown that levels of the mitochondrial biogenesis transcriptional 'master-regulator' PGC-1α in hippocampal tissues are reduced relative to control tissue, suggesting that disruption of PGC-1α-dependent pathways contributes to pathogenesis[92]. Collectively, this evidence points to a role for dysfunction of the mitochondrial-nuclear relationship in NDs.

Despite this, the analysis of mitochondrial-nuclear cross-talk at scale largely focuses on either on a small number of features, a small number of samples[93], *in vivo* work involving single gene knockdown[94] or indirectly analysing mitochondrial function by measuring metabolite output[95]. Larger studies that have looked at cross-talk in multiple tissues include a population-level analysis of expression quantitative trait loci (eQTLs) associated with the expression of mtDNA-encoded

genes, and a multi-tissue analysis of nDNA and mtDNA gene expression correlations[46,96]. These studies support the complexity and functional relevance of mitochondrial-nuclear relationships in the brain but lack CNS-specificity and analysis of potential processes and pathways most relevant to mitochondrial-nuclear coordination.

To further interrogate mitochondrial-nuclear regulation, this work takes a genome-wide view, surveying all potential nuclear-mitochondrial relationships and assaying the relationship using mtDNA-nDNA gene coexpression as a proxy for interaction. Coexpression is a useful metric with which to assess this relationship as it is well established that gene expression correlation can robustly predict gene function, and that genes which share biological processes are often co-regulated[97,98,99]. Coupled with this, the breadth of CNS tissues and sample abundance present in the GTEx project transcriptomic data provides an excellent opportunity for systematic evaluation of coexpression between the nuclear and mitochondrial genomes. The GTEx project is publicly available resource providing whole genome sequencing (WGS), RNA-Seq and whole exome sequencing (WES) data for many individuals across 53 body tissues[100,101]. Additionally, an AD case-control RNA-Seq dataset was leveraged to understand whether mitochondrial-nuclear coexpression changes between AD cases and healthy controls.

Specifically, this chapter generates and analyses mitochondrial-nuclear gene expression correlations, focuses on CNS tissues and gains power by leveraging RNA-Seq data from a large number of individuals across multiple studies. Robustness of mitochondrial-nuclear expression correlations is assessed by replicating results across two independent data sets. Finally, the contribution of the mitochondrial-nuclear relationship to neurodegeneration is interrogated by first analysing the association of causal ND genes with the mitochondrial genome, and then by analysing coexpression changes in an AD case-control dataset.

2.1.1 Aims

In light of this, this chapter's aims are two-fold. Firstly, it aims to set the groundwork for this thesis by characterising profiles of mitochondrial-nuclear correlations

in healthy CNS tissue. Secondly, it aims to understand whether mitochondrial-nuclear relationships are perturbed in ND, and whether these can be identified in CNS tissue.

2.2 Methods

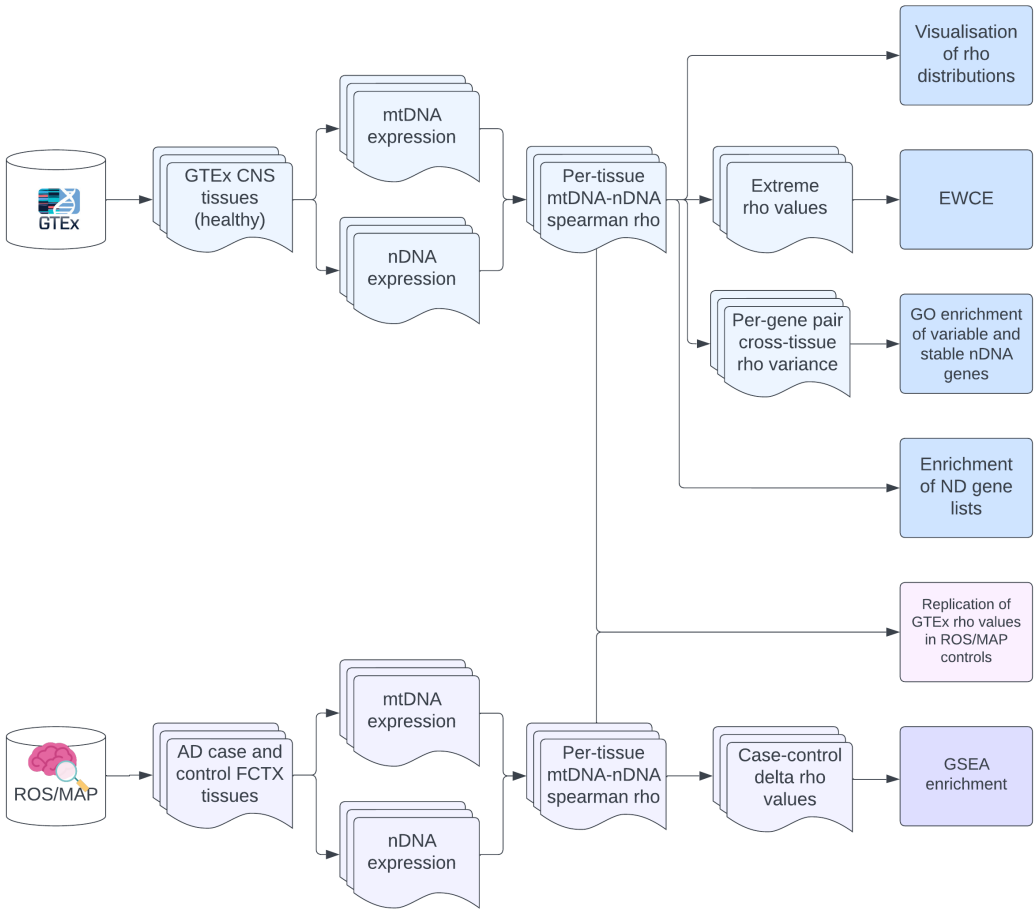

Figure 2.1: Workflow diagram to summarise the data and methods utilised in chapter 1.

2.2.1 Data

2.2.1.1 GTEx data

Raw RNA-Seq data from 12 histologically normal CNS regions were obtained from GTEX (V6p)[101]. Alignment and data processing was carried out by Aminah T Ali as per Ali et al. (2019). Briefly, adapter sequences, low-quality terminal bases and poly-A tails (>4) were trimmed and subsequently aligned to the 1000G GRCh37

reference genome using STAR. Strict filtering was applied to avoid misalignment of nuclear mitochondrial DNA segments (NUMTs) -derived sequences, and to retain only properly paired and uniquely mapped reads. Post-mapping processing included exclusion of samples with: $<$10K reads mapping to the mitochondrial genome, $<$5 million total mapped reads, $>$30% of reads mapping to intergenic regions, $>$1% total mismatches or $>$30% reads mapping to ribosomal RNA using custom scripts as well as RNAseQC. HTseq was used to quantify transcripts, before converting raw counts to transcript per million (TPM) values, therefore normalising for library size and gene length, using version 19 of the Gencode gene annotation. The final per-region sample numbers are shown in table 2.1.

GTEx tissue	Sample number (N)
Anterior cingulate cortex	80
Caudate basal ganglia	111
Cortex	104
Cerebellum	109
Cerebellar hemisphere	96
Frontal cortex	96
Hippocampus	88
Hypothalamus	85
Nucleus accumbens basal ganglia	100
Putamen basal ganglia	86
Spinal cord (cervical c1)	58
Substantia nigra	60

Table 2.1: GTEx (V6p) CNS per-tissue RNA-Seq sample numbers.

2.2.1.2 ROSMAP data

The ROSMAP dataset is composed of dorsolateral prefrontal cortex samples derived from autopsied individuals from the Religious Orders Study (ROS) and the Rush Memory and Aging Project (MAP)[102]. Data were obtained through application to the data access committee, permitting access to pre-mapped fragments per kilobase of transcript per million (FPKM) data (for sequencing, mapping and QC details see the Synapse Knowledge Portal, ID: syn3388564). Each ROSMAP sample is associated with a cognitive diagnosis. Samples included in the analyses were those labelled 'AD' (n=254) and 'no cognitive impairment' (n=201), referred to as 'case'

and 'control', respectively. Samples with missing metadata or had duplicates were removed, reducing the number of cases to 251. Prior to further processing, FPKMs were converted to TPMs.

2.2.2 Generating mitochondrial-nuclear correlation matrices

For both datasets, the same custom pipeline was applied to generate mitochondrial-nuclear gene expression correlation matrices from gene counts. First, TPM matrices were filtered, retaining the set of genes with non-zero expression across all samples and all CNS regions. TPMs were then log10 and median normalised. Expression outliers, defined as TPM values three interquartile ranges below the lower quartile or above the upper quartile for a gene, were masked.

Following this, multiple linear regression was applied to regress out covariates (see section 2.2.2.1 for covariate selection steps). TPM values were included as predictor variables and covariates as response variables in a linear model. The model was fitted, generating predicted TPMs, then residuals were calculated by subtracting predicted from observed, yielding residual TPMs. To generate mitochondrial-nuclear correlation matrices, Spearman correlation coefficients were calculated between protein-coding mtDNA-encoded genes (13) and nDNA-encoded genes (for GTEx: 15,001 genes expressed across all CNS tissues; for ROSMAP all nuclear genes expressed).

2.2.2.1 Selection of covariates

Covariates for data correction were selected by correlating axes of variation with known covariates. Known covariates were selected from the available metadata for being: a key attribute of the donor (i.e. age, sex), a key attribute of the sample, or an attribute likely to produce a batch effect. This selection was informed by exploratory analyses, and by established covariate selection strategies in the literature. To determine axes of variation in the expression data, principal component analysis (PCA) was performed on the expression matrices. Spearman correlations between the largest axes of variation (first 10 principle components, capturing 98.41% of the variation for GTEx and 99.43% for ROSMAP) and known covariates were cal-

culated (Figure 2.2). Covariates that correlated highly with large axes of variation were selected. For ROSMAP, this was: PMI (post-mortem interval), RIN (RNA integrity number), library batch, race, sex, study (whether subject was derived from the ROS or MAP cohorts of the dataset), age at death, age at last visit. For GTEx, this was: RIN, four batch variables (type of nucleic acid isolation batch, nucleic acid isolation batch ID, genotype or expression batch ID, date of genotype or expression batch), centre (location of tissue collection and processing), age, gender and cause of death.

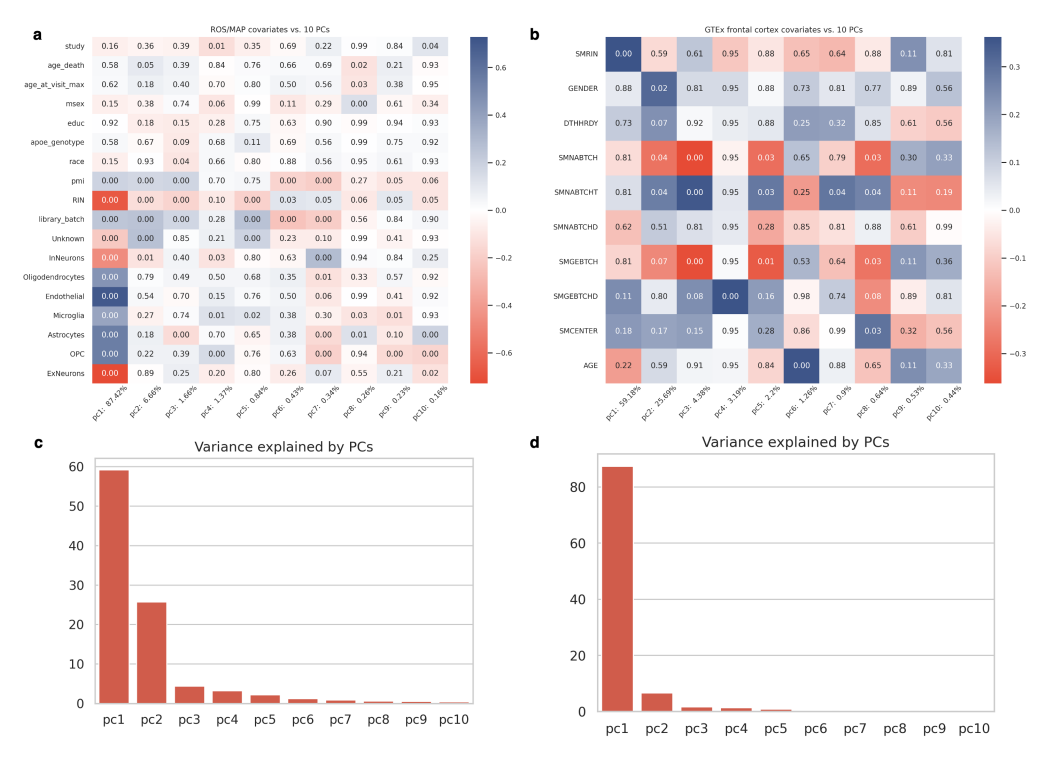

Figure 2.2: Steps taken to decided on the covariate correction strategy for ROSMAP and GTEx project -derived gene expression data. Here, colour bars represent the Spearman correlation coefficient and tile annotations give the P-value. A. Heatmap to show Spearman correlations between GTEx frontal cortex expression principle components (10) and known sample attributes (10) supplied by the GTEx project. B. Heatmap to show Spearman correlations between ROSMAP expression principle components (10) and known sample attributes (20) supplied by the ROSMAP project. C. Percent variance in the GTEx expression data explained by principle component 1-10. D. Percent variance in the ROSMAP expression data explained by principle component 1-10.

2.2.3 Analysing mitochondrial-nuclear correlation variance across CNS regions

To understand the extent to which the relationships between mitochondrial-nuclear gene pairs vary across the CNS, 12 GTEx CNS regions were leveraged, calculating the cross-CNS variance of correlation coefficients for each mitochondrial-nuclear gene pair. The variance of the 12 coefficients was then calculated as a measure of variation in the relationship between the expression of these two genes across CNS regions. This was repeated for all mitochondrial-nuclear gene pairs. Nuclear genes expressed in all 12 CNS regions were used, equating to 15,001 nuclear genes and 195,013 mitochondrial-nuclear pairs. To reduce redundancy of the dataset, aggregation of mtDNA-encoded genes was performed, the intuition being that the correlation of a nuclear gene with the 13 mtDNA-encoded genes was found to be largely consistent. Figure 2.3 shows that expression of each mtDNA-encoded gene is consistent across GTEx CNS regions (a), correlation of expression between mtDNA-encoded genes was high except for in the 'Cortex' tissue (b, c) and variance across the 13 mitochondrial-nuclear correlations for each nuclear gene was low (d). Thus the median cross-CNS variance of 13 mtDNA-encoded genes was taken as the representative value for each nuclear gene.

To determine processes enriched in gene pairs in different variance groups - i.e. those that vary across CNS tissues and those that do not - four gene sets were defined. The 'high variance set' (highest 5% of variances, n=750), and the 'low variance set' (lowest 5% of variances, n=750). These two groups were then further split into positive and negative sub-groups, dependent on the majority correlation directionality. This yielded the following gene sets: high variance positive, n=605; high variance negative, n=145; low variance positive, n=363; low variance negative, n=387.

To determine the processes and pathways enriched in these gene sets, the R package gProfiler2 was used[103]. Enrichments were tested against a custom background of genes expressed in all GTEx CNS regions (n=15,001). The queries were ordered by correlation magnitude, and for multiple test correction, the 'g:SCS'

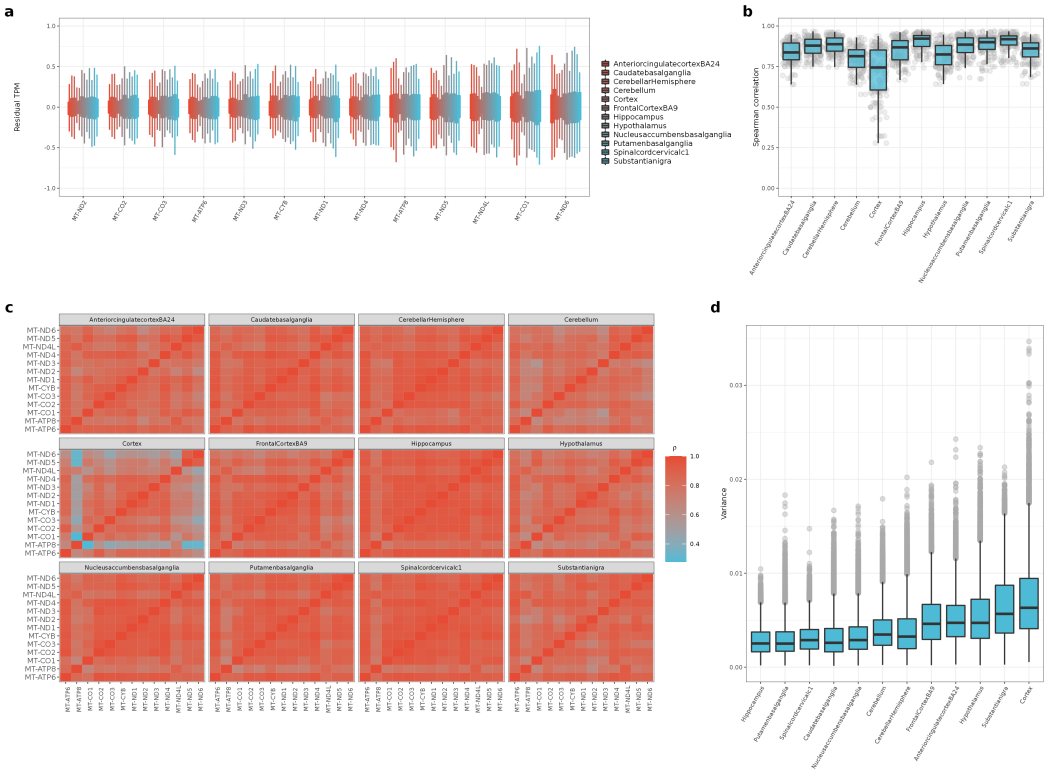

Figure 2.3: Plots to show mtDNA gene expression homogeneity within brain regions. A. Per-mtDNA gene TPM residuals, coloured by GTEx brain region. B. Per-region mtDNA-mtDNA gene Spearman ρ distributions. C. Per-region mtDNA-mtDNA gene correlation Spearman ρ values. D. Distribution of nuclear-mitochondrial correlation variances, with variances calculated across 13 mtDNA genes for each nDNA gene.

method was applied. To obtain a more granular ontology analysis of the synaptic enrichment observed in the high variance negative group, this list was used as input to the online tool SynGO[104]. The same background list was used for SynGO as for gProfiler2.

2.2.4 EWCE analysis

Expression Weighted cell type Enrichment (EWCE)[105] was used to determine whether nuclear gene sets had higher expression within particular CNS cell types than would be expected by chance. EWCE leverages single-nuclear RNA-seq (snRNA-Seq) data in the form of specificity matrices. Specificity matrices give, for each gene and each cell type, the expression specificity a gene has in a cell type compared with all other cell types. Using this information, EWCE statistically eval-

uates whether genes in a target list have a higher cell type specificity than would be expected by chance (i.e. than the random distributions drawn from the background).

Inputs to EWCE were target gene lists, a background gene set and a specificity matrix. Aggregation over mtDNA genes was then performed to obtain a single consensus ranking for each nuclear gene. The target gene lists used were generated by ranking mitochondrial-nuclear gene correlation values for each GTEx CNS region with the largest positive and negative values ranked separately. For each CNS region, the top 5% of positively correlated nuclear genes and top 5% of negatively correlated nuclear genes were then taken as region-specific target gene sets. The numbers of genes per region are given in table 2.2. The background gene set was genes expressed in all GTEx CNS regions (n=15,001). The specificity matrices were generated by Regina H Reynolds as in Mencacci et al., (2020) by estimating the specificity of each gene to each cell type. The specificity score represents the proportion of the total expression of a gene found in one cell type compared to all cell types. Data used to generate specificity matrices for this work were derived from two brain snRNA-seq experiments. (1) The Allen Brain Atlas: a dataset comprising 15,928 nuclei from the middle temporal gyrus of 8 human tissue donors ranging in age from 24 to 66 years[106]. (2) Habib et al., (2017): a dataset comprising 19,550 nuclei from the hippocampus (4 samples) and prefrontal cortex (3 samples) from five donors.

The EWCE analysis was run with 10,000 bootstrap lists. Transcript length and GC-content biases were controlled for by selecting bootstrap lists with equivalent properties to the target list. P-values were corrected for multiple testing (using the Benjamini-Hochberg method) over all cell types and gene lists tested. The analysis was performed using the following major cell type classes: ‘GABAergic’, ‘glutamatergic’, ‘astrocyte’, ‘microglia’, ‘oligodendrocyte’, ‘endothelial cell’.

2.2.5 Cell type correction analysis

To further evaluate the contribution of cell type to the heterogeneity of mitochondrial-nuclear correlation distributions in the CNS, published deconvolution-derived GTEx cell type proportion estimates from Donovan et al., (2020) were

GTEx tissue	No. of nuclear genes (+ρ and -ρ)
Anterior cingulate cortex	470
Caudate basal ganglia	503
Cortex	340
Cerebellum	277
Cerebellar hemisphere	399
Frontal cortex	533
Hippocampus	426
Hypothalamus	268
Nucleus accumbens basal ganglia	487
Putamen basal ganglia	492
Spinal cord (cervical c1)	336
Substantia nigra	312

Table 2.2: Numbers of unique nuclear genes in the top 5% positive and top 5% negative mitochondrial-nuclear gene Spearman ρ, for each GTEx CNS tissue.

used[107]. The aim was to understand whether correcting the GTEx expression data for the effect of cell type proportions would result in more homogeneous mitochondrial-nuclear correlation profiles across the CNS. To do this, the effect on mitochondrial-nuclear correlation distributions was compared between two correction strategies, one which does not correct for cell type, and one which does. These were defined as follows: (i) standard correction (covariates: RIN, four batch variables (type of nucleic acid isolation batch, nucleic acid isolation batch ID, genotype or expression batch ID, date of genotype or expression batch), centre, age, gender and cause of death), (ii) standard & cell type correction (covariates: as in (i) as well as cell type proportions for the following cell types: astrocyte of the cerebral cortex, Bergmann glial cell, brain pericyte, endothelial cell, neuron, oligodendrocyte and oligodendrocyte precursor cell). The sample set used for the calculation of mitochondrial-nuclear correlations differed slightly to those used for the other analyses presented in the chapter due to availability of cell type proportion data (see table 2.3. Following filtration for these samples, the pipeline as described in 2.2.2, was run, producing mitochondrial-nuclear correlation matrices for both correction strategies.

This analysis relied on the assumption that the cell type proportions were rep-

GTEx tissue	Sample number (N)
Anterior cingulate cortex	65
Caudate basal ganglia	86
Cortex	78
Cerebellum	89
Cerebellar hemisphere	69
Frontal cortex	76
Hippocampus	67
Hypothalamus	62
Nucleus accumbens basal ganglia	74
Putamen basal ganglia	64
Spinal cord (cervical c1)	49
Substantia nigra	43

Table 2.3: Sample numbers input to GTEx cell type correction analysis, resulting from samples available for each GTEx tissue which also had available cell type proportion data (as derived from Donovan et al., (2020)).

resentative of the cellular composition of GTEx CNS regions. However, several factors affected the accuracy of the proportion estimates. Firstly, they were derived using murine brain scRNA-Seq data, representing a species mismatch. Secondly, technical factors such as the dissection protocol (excision order and resulting RNA degradation), the size of each target region and the accuracy of tissue excision. In the Donovan et al., (2020) scRNA-Seq data, cell type proportions were assigned to the GTEx spinal cord tissue, which was not in fact dissected from the mice, as per the author's dissection protocol[107]. However, proportions for the spinal cord were indeed present in the cell type proportion data. As such, these proportions cannot be considered to accurately represent the cell type composition of the spinal cord. Considering these factors, a subset of GTEx CNS regions that were most accurately represented by the available cell type proportion estimates were selected. To robustly select these, 20 axes of variation were calculated from GTEx CNS RNAseq data (using PEER v1.052[108]) on all samples, obtaining 20 PEER factors for each region. These were correlated with the cell type proportion estimates to understand whether the cell type proportions represented the cell type composition of each GTEx CNS region. The underlying logic here was that larger axes of variation would correlate strongly with the cell type proportion estimates, but only if those

estimates accurately represented the tissue.

Five CNS regions (anterior cingulate cortex, cortex, frontal cortex, hippocampus, caudate basal ganglia) were identified for which top PEER factors (explaining most variance in the data) were correlated with cell type proportions. In these regions there were fewer spurious correlations with cell types across the 20 PEER factors, and PEER factors explaining most variance (PEER 1 and 2) were highly correlated ($\rho \geq 0.6$) with cell types, indicating better representation of these regions by the cell type proportion data. To obtain a measure for cross-CNS variation of mitochondrial-nuclear relationships, variances were calculated for each mitochondrial-nuclear gene pair across these five regions. This was done for correlation values produced by both correction strategies. Finally, to understand whether the cell type influenced cross-CNS variation in mitochondrial-nuclear correlation distributions, a two-sample Wilcoxon signed rank test was performed. The null hypothesis was that the true location shift from standard to standard and cell type correction -derived distributions was <0 (i.e. a negative shift in variance, closer to a median of zero in the latter). One-sample Wilcoxon tests were also performed for each correction strategy to test the null hypothesis that the median of the distribution of variances was equal to zero.

2.2.6 Testing disease-relevant gene lists against a random background

The aim of this analysis was to determine whether specific disease-relevant gene sets had more extreme distributions of mitochondrial-nuclear gene expression correlations than a random, equally sized, set of genes. To this end, four gene sets were analysed (sets 1, 2, 5 and 6 below). The analysis was subsequently expanded to incorporate seven gene sets in order to validate emerging findings and further test the hypothesis with smaller, more specialised sub-lists of disease genes. The total catalogue of the seven sets used is as follows: (1) A set of 35 AD-associated genes of interest were derived from a recent AD GWAS[109]. This study analysed SNPs in clinically diagnosed cases (71.88K) and controls (383.378K), identifying >20 AD-associated loci. (2) A set of 62 PD-associated genes of interest were selected

on the basis of eQTL data from a recent PD GWAS[65]. This study analysed 7.8M SNPs in 37.7K cases, 18.6K UK Biobank proxy-cases, and 1.4M controls, identifying 90 signals at genome-wide significance. The Genomics England PanelApp tool gives sets of clinically curated genes associated with disease through rare high effect variants[110]. The following panels were downloaded from this resource: (3) Early onset dementia (28 genes). (4) PD and complex PD list (43 genes). This panel contains genes associated with early onset and familial Parkinson's disease as well as complex Parkinsonism. (5) Adult onset ND disorders (110 genes). This panel is a super-set, including the early onset dementia and PD PanelApp panels as well as genes from other ND-related panels wherein variants are known to cause ND. (6) Intracerebral calcification disorders (21 genes) used as a negative control because the pathogenesis of these disorders is distinct from AD and PD. (7) A set of genes curated by OMIM (Online Mendelian Inheritance in Man) including genes associated with PD phenotypes (24 genes, OMIM accession: #PS168600).

For each GTEx CNS region, r, and each gene set, l, the median mitochondrial-nuclear correlation value of l for r was calculated. The distribution of mitochondrial-nuclear pairs was inclusive of all mitochondrial correlations for each nuclear gene, making the size of the distribution (length of l)*13. To generate empirical distributions, a random sample of nuclear genes of matching biotype and length, l, was selected from the set of genes expressed in all GTEx CNS regions (15,001) and all correlations with mtDNA genes were included.

A two-tailed test was carried out to determine whether l had a more extreme median mitochondrial-nuclear correlation value than could be expected by chance. To this end, the median of l was compared to the medians of 10,000 randomly selected gene sets. P-values were calculated as follows, where k is the number of randomly selected sets, and n is the number of correlations more extreme than the median of l:

$$P = (k \pm n)/k$$

A series of significance thresholds of increasing stringency were calculated to reflect the number of tests performed, taking into consideration the num-

ber of tissues and the number of gene sets analysed. The significance of results was assessed against the following Bonferroni multiple-test corrected P value thresholds: 0.05/12<P<0.05; 0.05/12*(number of gene sets)<P<0.05/12; P<0.05/12*(number of gene sets).

2.2.7 Case-control analysis of ROSMAP data

To identify mitochondrial-nuclear gene pairs that are modulated in disease states, the ROSMAP case-control AD dataset was used. Due to cell type proportion changes related to disease pathogenesis in AD brain tissue, correction for cell type proportion was carried out (additional to the covariates listed in section 2.2.2.1) using deconvolution-derived cell type proportions[111,112,113]. The cell type proportion distributions for the case and control ROSMAP data are shown in figure 2.4. To quantify changes in mitochondrial-nuclear coexpression, aggregation over mtDNA-encoded genes was carried out for the case and control data separately by taking the median Spearman's ρ value for each nuclear gene. The difference between these values was then calculated ($\rho_{control}$ - ρ_{case}) for each gene pair, giving case-control delta values, $\Delta\rho$.

To identify pathways enriched in high $\Delta\rho$ values (i.e. pairs with large case-control disparities), the GSEA method was applied using the fGSEA R package[114]. The inputs into fGSEA were gene lists ranked by $\Delta\rho$ and split by directionality. With a separate positive and negative correlation list, the sign of the $\Delta\rho$ in each case relates to whether a gene pair's correlation magnitude has increased or decreased in case in comparison to control. As such, any enrichments are interpretable as being related to case-control shifts. The fGSEA parameters used were as follows: 'GO' as the annotation source; minimum and maximum size of terms set to 15 and 2000 respectively. fGSEA was run using the fgseaMultilevel function and output was visualised using the plotGseaTable function.

2.3 Results

Since mitochondrial processes are important in CNS tissue and their perturbation is thought to have a role in several NDs, the aim was to identify whether relationships

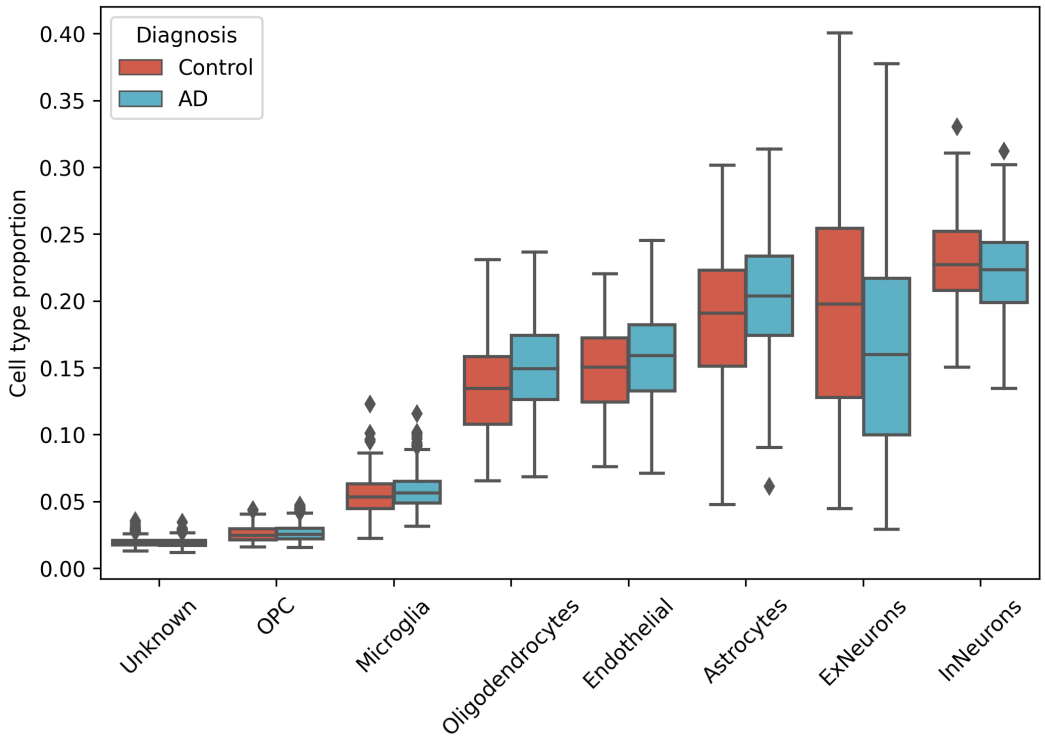

Figure 2.4: Distributions of ROSMAP cell type proportions derived by Donovan et al., (2020) for control samples (blue), and AD case samples (red).

between expression levels of mtDNA and nDNA -encoded genes vary across CNS regions, cell types and ND status. To do this, Spearman correlation coefficients were calculated between all mtDNA and nDNA -encoded gene pairs, after regressing out covariates (see section 2.2.2.1). RNA-Seq was leveraged from 12 CNS tissues derived from the GTEx project for analyses in healthy tissue, and frontal cortex tissue from the ROSMAP project AD dataset for analyses in a case-control paradigm.

2.3.1 Mitochondrial-nuclear correlations across the human CNS

In order to investigate correlations in mitochondrial-nuclear gene expression across all CNS regions, Spearman correlation coefficients were calculated for each pair of nDNA and mtDNA -encoded genes (15,001 and 13 genes respectively, making a total of 195,013 comparisons) in each of the 12 GTEx CNS regions. Distributions of the correlation values for each CNS region were visualised as density plots to facilitate cross-CNS comparison (Figure 2.5a). CNS regions were ob-

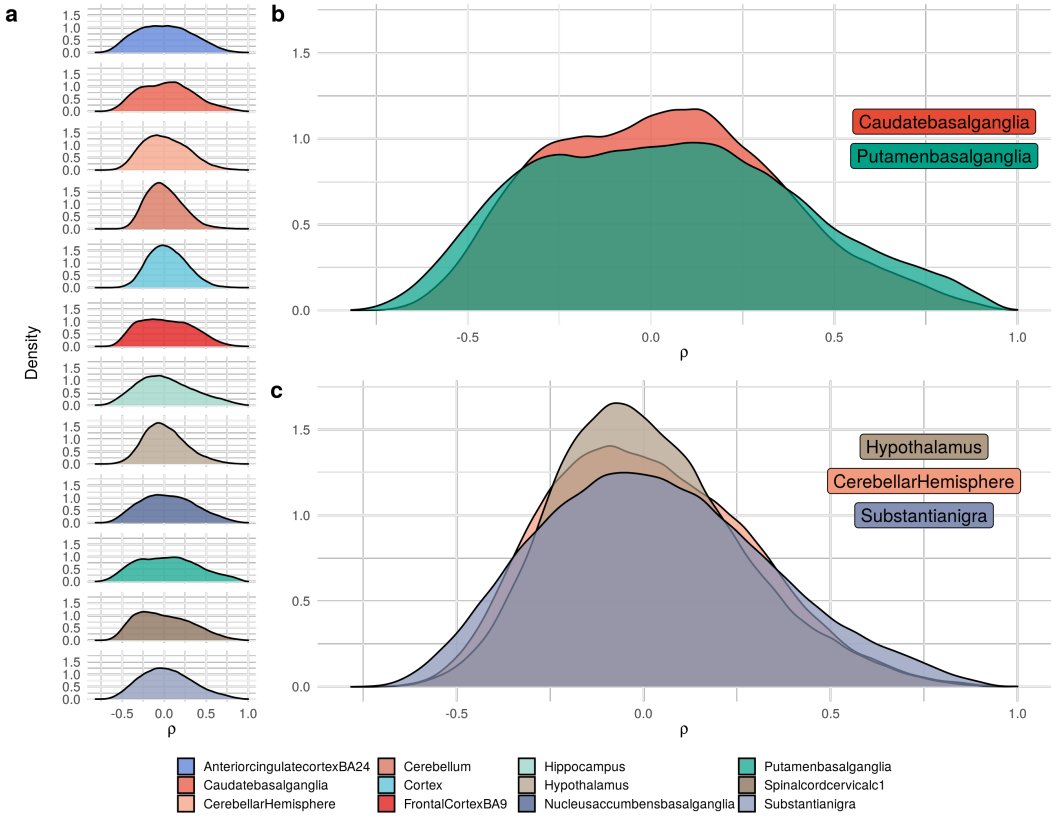

Figure 2.5: Distributions of mitochondrial-nuclear expression correlation coefficients (Spearman ρ) for each GTEx CNS region. A. Mitochondrial-nuclear ρ distributions for 12 GTEx CNS regions. B. Panel to show ρ distributions of the putamen basal ganglia and caudate basal ganglia regions. C. Panel to show distributions of the cerebellar hemisphere, hypothalamus, and substantia nigra regions.

served to have distinct and variable mitochondrial-nuclear correlation distributions. While some regions showed Gaussian-like distributions (cerebellar hemisphere, hypothalamus, substantia nigra) (Figure 2.5c), others showed dispersed distributions, containing more high magnitude relationships, and fewer neutral correlations (caudate basal ganglia, putamen basal ganglia) (Figure 2.5b). Qualitative analysis revealed mitochondrial-nuclear distribution similarity within GTEx CNS tissues derived from the same broad regional classification (fore-brain, mid-brain and hind-brain). To confirm this quantitatively, unsupervised Euclidean clustering of regional correlation coefficients across all CNS tissues was performed (Figure 2.6). This identified biologically meaningful clusters, whereby cortical regions and distinct regions of the basal ganglia (putamen, nucleus accumbens and caudate) were grouped

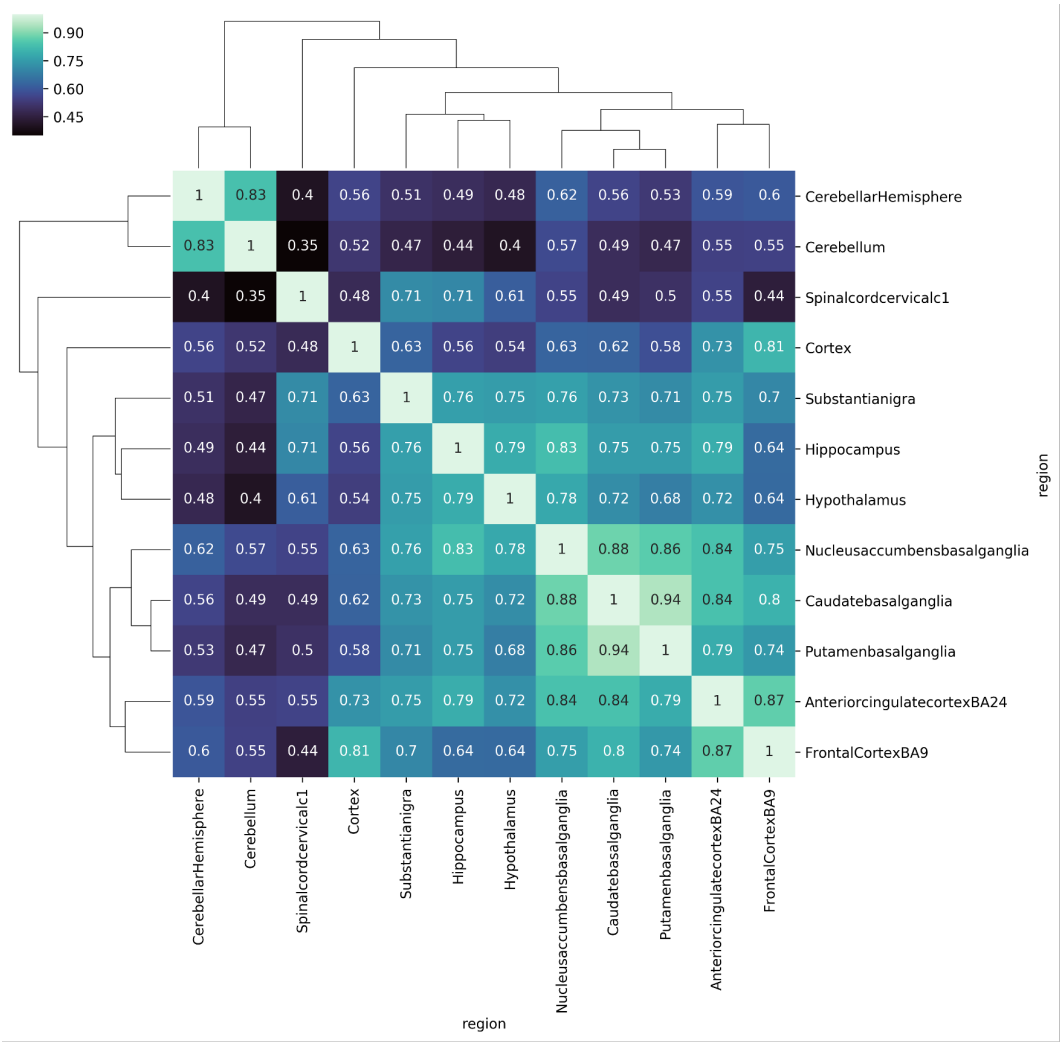

Figure 2.6: Euclidean clustering of 12 GTEx CNS regions, where clustering was performed across tissues based on tissue-tissue correlation coefficients (indicated on each tile, Spearman ρ).

together, which appears to reflect functional specialisation in the human brain.

2.3.2 The contribution of cell type composition to mitochondrial-nuclear correlations

To understand the patterns observed in figure 2.5, the following hypothesis was formed: regional differences in cell type composition contribute to regional differences in mitochondrial-nuclear correlation profiles. To test this, it was considered whether cell type markers were enriched at the positive and negative extremes of the correlation coefficient distributions. This analysis was performed for each GTEx

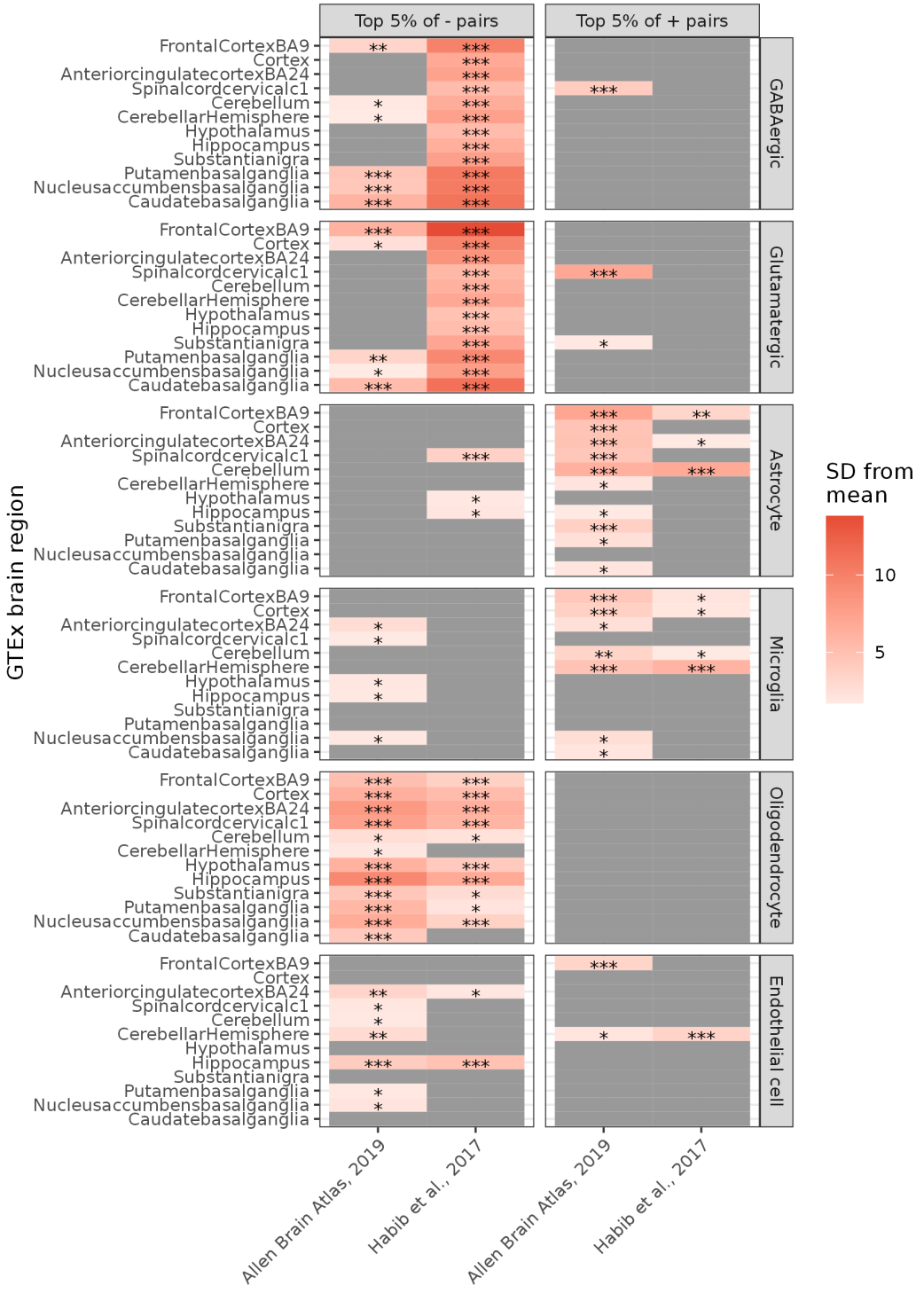

Figure 2.7: EWCE-derived cell type enrichments for 12 GTEx CNS regions. The left-hand y-axis refers to the GTEx CNS region, while the right-hand facet labelling refers to the cell type. For each cell type in each region, the metric for enrichment is shown as the number of standard deviations from the bootstrapped mean (SD from mean, indicated by the colour bar). The x-axis indicates which scRNA-seq dataset the underlying cell type specificity matrix was derived from. For each association, the following asterisks are overlaid to indicate the multiple test correction threshold passed: *0.05/12 <P< 0.05; **0.05/12*6<P< 0.05/12; ***P<0.05/12*6.

CNS region using the Expression Weighted Cell Type Enrichment (EWCE) method, which tests whether a given set of genes is expressed more highly in a cell type of interest than might be expected by chance[105]. Cell type specificity data were derived from two human brain snRNA-Seq experiments, the first of which used middle temporal gyrus nuclei[115], and the second used hippocampus and prefrontal cortex nuclei[116]. The input to this method was nDNA-encoded genes derived from gene pairs in the highest 5% of positive correlations and highest 5% of negative correlations for each region.

Genes with a high specificity for neuronal cell types (GABAergic and glutamatergic) were significantly enriched ($P<0.05$, Bonferroni-corrected for regions and gene sets) in negative mitochondrial-nuclear gene pairs across CNS regions (Figure 2.7). In contrast, genes with a high specificity for non-neuronal cell types (astrocytes, microglia) were significantly enriched in positive mitochondrial-nuclear gene pairs ($P<0.05$ in 6/12 regions for astrocytes; $P<0.05$ in 5/12 regions for microglia, Bonferroni-corrected for regions and gene sets), the exception to this being oligodendrocytes (Figure 2.7). A strong cross-CNS signal for oligodendrocyte marker enrichment was observed in negatively correlated pairs ($P<0.05$ in 10/12 regions, Bonferroni-corrected for regions and gene sets), coupled with no significant enrichment detected in positively correlated pairs. For astrocytes and microglia, a trend towards marker enrichment in positive pairs over negative pairs across the CNS was observed. Reassuringly, related regions displayed similar cell type enrichment profiles, indicative of biological functionality being reflected in these enrichments. For example, GTEx-defined[100] technical sample replicates (the cortex and frontal cortex, and cerebellum and cerebellar hemisphere) as well as regions closely biologically associated such as the basal ganglia (putamen, nucleus accumbens and caudate), demonstrated consistent patterns of cell type enrichment.

To further test the posited hypothesis, published cell type proportion estimates[107] were leveraged to determine whether correcting GTEx expression data for the effect of cell type proportions would result in more homogeneous cross-CNS mitochondrial-nuclear correlation profiles. To this end, five GTEx regions (see sec-

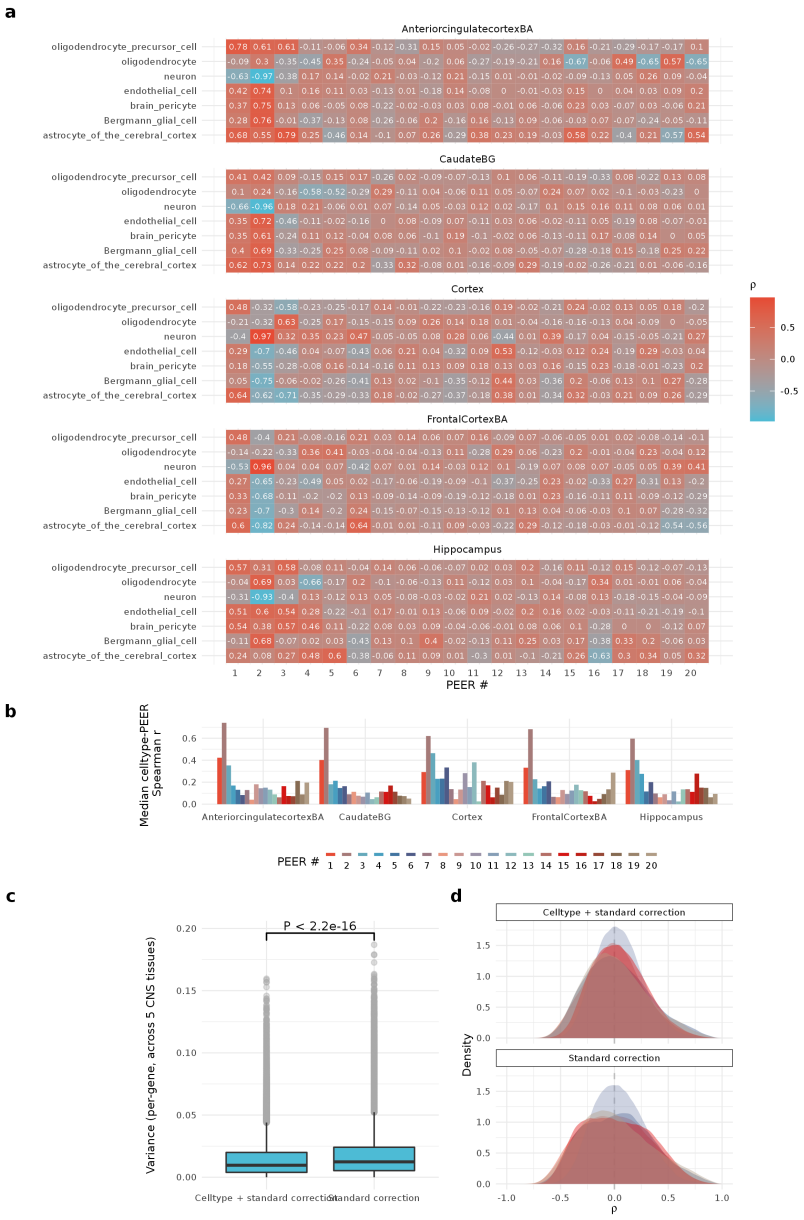

Figure 2.8: Demonstrating the effect of cell type correction on mitochondrial-nuclear correlation distributions in GTEx data. A. Heatmap to show correlations between PEER factors (generated from TPM values) and estimated cell type proportions for five GTEx CNS regions selected to be best represented by these estimated cell type proportions. Annotations correspond to Spearman's ρ values. B. A barplot to summarise part A, the median PEER cell type correlation value for each PEER factor is shown for the same five CNS regions. C. Distribution of gene-pair (195000 mitochondrial-nuclear pairs) variances across the five GTEx CNS regions for both correction strategies (standard and standard-cell type). A Wilcoxon signed rank test was carried out to test the null hypothesis that the median of the cell type -standard corrected correlation data was not less than that of the standard corrected data (P<2.2e-16). Boxplots display the median, upper quartile (Q3) and lower quartile (Q1), with whiskers extending to Q3+1.5*IQR and Q1-1.5*IQR. Diamond points represent outliers above Q3+1.5*IQR or below Q1–1.5*IQR. D. Density plot for the five selected CNS tissues, faceted by correction, showing the distribution of nuclear-mitochondrial correlation values under each correction strategy.

tion 2.2.5) were included for which the cell type proportions were determined to be most representative (Figure 2.8a and b). Distributions of cross-regional Spearman correlation variances per mitochondrial-nuclear gene pair were compared with and without correction for cell type proportions. Applying this approach, distributions of variances were found to be significantly different to each other (two-sample Wilcoxon signed rank test, P<2.2e-16), but the medians of both distributions were also significantly higher than 0 (one-sample Wilcoxon signed rank test, P<2.2e-16 for mitochondrial-nuclear distributions derived from both correction strategies) (Figure 2.8c and d). Thus, cell type proportion was been found to be a modulator of cross-CNS variation in mitochondrial-nuclear correlations. However, it is important to note that regional specialisations still existed after correcting for cell type proportions.

2.3.3 Cross-CNS variability of mitochondrial-nuclear correlations

Having established the importance of cell type composition in driving variation mitochondrial-nuclear correlation profiles variation across the CNS, the aim was to identify biological processes associated with this variation. To find out which mitochondrial-nuclear gene pairs varied across the CNS, the variance of correlations for each mitochondrial-nuclear gene pair across the CNS was calculated. To reduce redundancy, aggregation of mtDNA-encoded genes was performed, taking the median cross-CNS variance of 13 mtDNA genes as the representative value for each nDNA gene (validity of this demonstrated in Figure 2.3, where homogeneity of mitochondrial-nuclear correlation across mtDNA genes can be observed). A correlation directionality was then assigned to each pair (see example in Figure 2.9a). Using this methodology, four gene sets were defined: (1) 'high variance positive': top 5% nDNA genes with the most variable relationships with the mitochondrial genome across brain regions (N=605); (2) 'high variance negative': top 5% nuclear genes with the most variable relationships with the mitochondrial genome across brain regions (N=145); (3) 'low variance positive' (N=387); (4) 'low variance negative' (N=387). These gene sets were used as input for the gene ontology enrichment

tool gProfiler2 to derive enriched pathways[103].

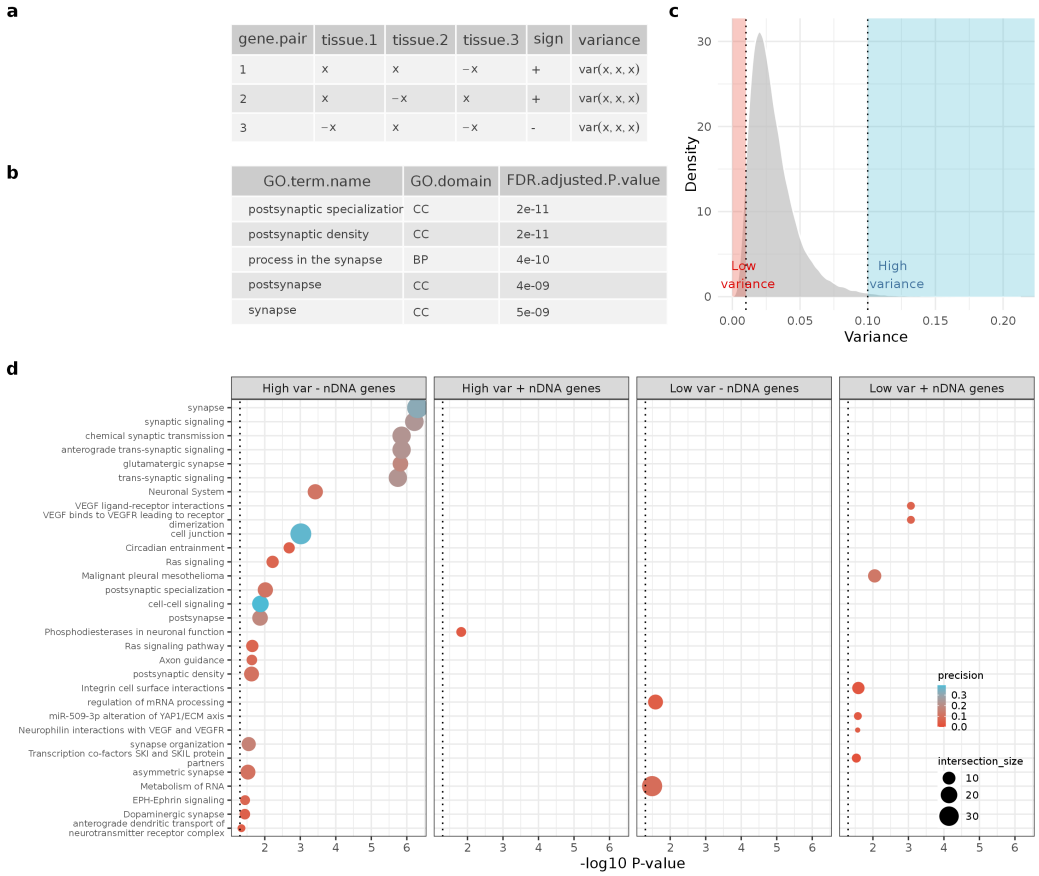

Figure 2.9: Visualisation of cross-CNS gene pair correlation variances in GTEx data and the processes enriched in four variance-defined gene sets. A. Schematic to visualise generation of cross-CNS variances. For each mitochondrial-nuclear gene pair, a variance is taken of its per-tissue Spearman's ρ values. The pair is assigned a directionality (sign) based on the majority directionality of its ρ values. B. Density plot of the distribution of cross-CNS mitochondrial-nuclear gene pair variances. The left-hand dotted line enclosing the shaded red area is the cut-off for 'low variance' gene pairs, the right-hand dotted line enclosing the blue shaded area is the cut-off for 'high variance' gene pairs. C. SynGO (synaptic gene ontology) output showing the top five enrichments for the high variance negative nuclear gene set. P-values are FDR-adjusted. D. gProfiler2-derived enrichments for four nuclear gene sets: high variance negative, high variance positive, low variance negative and low variance positive. The dotted line indicates a 5% significance cut-off. P-values were corrected using the gProfiler g:SCS method, optimised for enrichment analysis P value correction.

Overall, the distribution of variances was highly skewed towards zero, demonstrating that the vast majority of mitochondrial-nuclear pairs are stably correlated across all CNS regions (Figure 2.9c). In gene pairs that showed consistency across

CNS regions, enrichment was observed for VEGF ligand-receptor interactions in the positive correlation set (set 3 above, P=8.12e-04, corrected for multiple tests), whereas RNA processing (set 4 above, P=7.72e-3, corrected for multiple tests) was enriched in the negative correlation set (Figure 2.9d). Amongst the nuclear genes with the most variable relationships to the mitochondrial genome across brain regions, enrichment of phosphodiesterases in neuronal function as the only significant term for the positive (set 1) and synaptic terms in the negative set (set 2), with the most significant term being glutamatergic synapse (P=1.42e-06, corrected for multiple tests) (Figure 2.9d). To explore this enrichment further, SynGO[104], a specialist synapse ontology enrichment tool, was leveraged and found significant enrichment in the high variance negative list only. This set was highly significantly enriched for postsynaptic terms (P=3.4558e-20, FDR-corrected) with 3/5 of the most significant terms relating to this structure (Figure 2.9b). Of the 28 significant terms, 13 related to 'postsynaptic' structures or processes and 5 related to 'presynaptic' (see table A.1 in Appendix A for all significant terms). Overall, this analysis identified sub-cellular specificity in mitochondrial-nuclear correlations across the CNS. More specifically, variable mitochondrial-nuclear relationships highlighted genes associated with postsynaptic processes.

2.3.4 Replicating correlations in independent data

To determine whether the patterns of mitochondrial-nuclear correlation observed in GTEx brain data were robust, neurological control samples from the ROSMAP dataset were utilised. Since ROSMAP data are derived from dorsolateral prefrontal cortex tissue, the findings were compared to those generated from the GTEx frontal cortex tissue only, utilising the same pipeline was used for the GTEx data.

Overall, Spearman's ρ values for all mitochondrial-nuclear gene pairs showed high correlations between GTEx and ROSMAP data (Spearman's ρ=0.59, P<2e-16, for 13,640 nDNA genes that were expressed in both datasets), highlighting the consistency of mitochondrial-nuclear relationships in CNS tissue (Figure 2.10a). Inspection of correlation distributions across the two datasets revealed greater similarity at high Spearman's ρ magnitudes, likely due to the greater certainty associated

with those correlations (Figure 2.10a). Next, the replicability of the top 5% (ranked by Spearman correlation magnitude) positively and negatively correlated gene pairs was analysed. This found that 817 nuclear genes were in the top 5% of negative pairs for both datasets, and 588 nuclear genes were found in the top 5% of positive pairs for both datasets (Figure 2.10b). As such, 36% (top 5% positive) and 52% (top 5% negative) of the GTEx-derived gene sets are composed of the same genes when derived from ROSMAP data.

Given these findings, replication analyses were subsequently extended to understand whether cell type -specific enrichments identified in GTEx frontal cortex were robust across datasets. Repeating the EWCE analysis (see section 2.2.4) using the top 5% positive and negative gene lists generated from the ROSMAP control data (Figure 2.10c), highly similar patterns of cell type enrichment can be observed across datasets. There was significant enrichment of genes with high neuronal specificity in negatively correlated mitochondrial-nuclear pairs ($P<0.05$, Bonferroni-corrected for regions and gene sets) (Figure 2.10c). Additionally, there was significant enrichment of genes with high specificity to astrocytes ($P<0.05$, Bonferroni-corrected for regions) and microglia ($P<0.05$, Bonferroni-corrected for regions and gene sets) amongst positively correlated mitochondrial-nuclear pairs. Enrichment of oligodendrocyte marker genes in negative pairs was also replicated in the ROSMAP frontal cortex data ($P<0.05$, Bonferroni-corrected for regions and gene sets) (Figure 2.10c). Thus, robust replication of EWCE cell type enrichments in the ROSMAP data was achieved. Specifically, neuronal enrichment in the negative mitochondrial-nuclear space, and glial enrichment in the positive space were found to be highly reproducible.

2.3.5 Examining mitochondrial-nuclear correlations for genes implicated in neurodegenerative disease

Given the robust nature of mitochondrial-nuclear relationships and their association with specific cell types in CNS tissue, the aim was to investigate whether genomic cross-talk is relevant to the aetiology of NDs. To this end, the analysis tested whether mitochondrial-nuclear correlation distributions for genes implicated

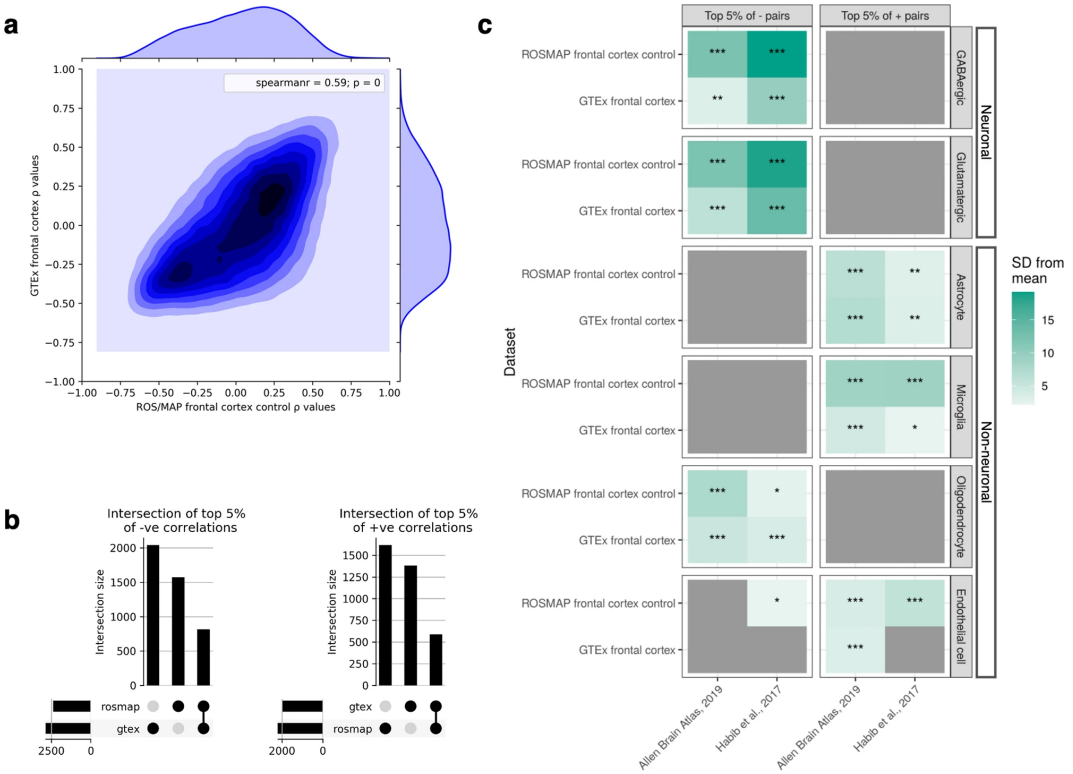

Figure 2.10: Replication of the mitochondrial-nuclear correlation values and cell type enrichments discovered in GTEx frontal cortex in an independent frontal cortex dataset (ROSMAP control samples). A. Density-contour plot to show all mitochondrial-nuclear gene pairs commonly expressed in both datasets (177,320). ROSMAP ρ values are plotted on the x-axis, and GTEx ρ values are plotted on the y-axis. The Spearman correlation for the overall bi-dataset correlation and corresponding P-value for the ρ statistic is given in the top right of the plot (Spearman's ρ=0.59, P<2e-16). B. Upset plots to show numbers of unique nuclear genes found in the top 5% positive (left-hand plot) and top 5% negative correlations in the two datasets, and the overlap size of these gene sets. 817 nuclear genes were found in the top 5% of negative pairs for both datasets, and 588 nuclear genes were found in the top 5% of positive pairs for both datasets. Thus, 52% and 36% of unique nuclear genes from negative and positive mitochondrial-nuclear pairs discovered in GTEx replicate in the ROSMAP control dataset. C. EWCE-derived cell type enrichments for GTEx frontal cortex and ROSMAP frontal cortex. The y-axis denotes the RNA-seq source. For each association, the following asterisks are overlaid to indicate the multiple test correction threshold passed: *0.05/12<P<0.05; **0.05/12*6<P<0.05/12; ***P<0.05/12*6.

in NDs were significantly different to distributions generated using random sets of genes (Figure 2.11).

Four gene sets were tested in the first instance: two sets derived from AD[109] and PD[65] GWA studies respectively (implicating genes through analyses of common variants), a gene set from the Genomics England PanelApp containing genes implicated in rare Mendelian forms of adult onset ND, and a second PanelApp list, intracerebral calcification disorders[110], as a negative control. These were largely independent sets, with minimal overlap in the genes included (for visualisation of gene set overlaps, see upset plot in Figure 2.12). Genes associated with AD through GWA studies were found to have mitochondrial-nuclear correlations which were nominally different (did not pass multiple test correction) from random gene sets in cortex (P=0.0206) and substantia nigra tissues (P=0.0273) (Figure 2.11a). Similarly, a nominally significant distribution shift was observed in hypothalamus tissue using the gene set implicated in sporadic PD (P=0.0163). In contrast to this, genes associated with adult onset ND displayed highly significant shifts in the majority of CNS regions ($P<0.05$, Bonferroni-corrected for regions and gene sets).

To test whether these findings were specific to a subset of NDs, mitochondrial-nuclear correlations among genes implicated in intracerebral calcification disorders (ICDs) were utilised as a negative control. Unlike AD and PD, ICD-induced neurodegeneration is caused by calcium deposition in the brain's vasculature or parenchyma. No significant difference was found between this gene set and empirical distributions in any CNS tissues.

The PanelApp adult onset ND gene set is an umbrella set, incorporating genes in the smaller and more specific 'early onset dementia' and 'PD and complex PD' PanelApp gene sets (for visualisation of overlaps, see upset plot in Figure 2.12b). As such, it was asked whether these more specific disease-related subsets also had significant relationships to the mitochondrial genome. The analysis was subsequently expanded to include these gene sets, and set more stringent significance cut-offs to consider the increased number of tests. Genes implicated in Mendelian forms of PD (PanelApp 'PD and complex PD') showed significant differences in mitochondrial-

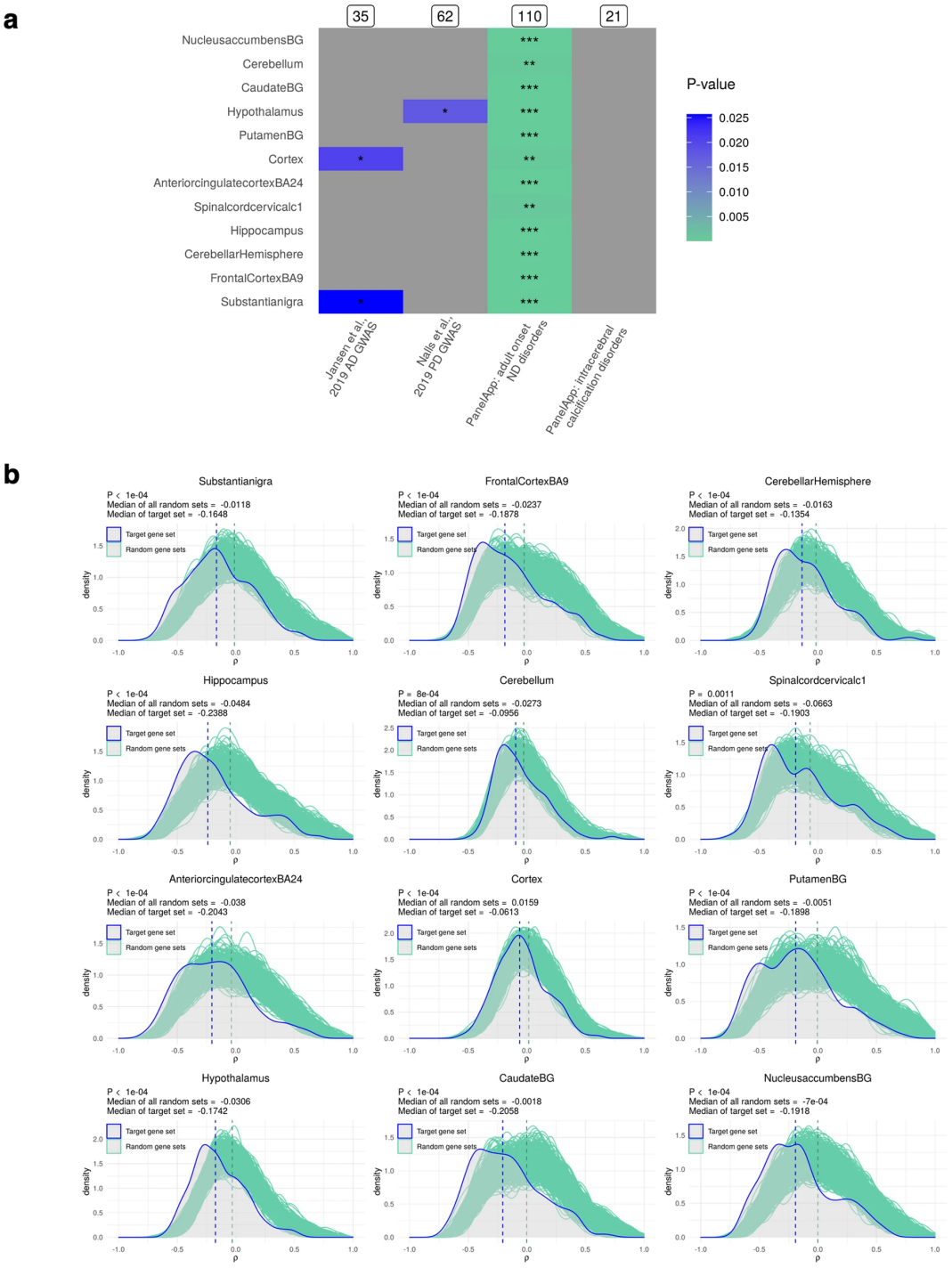

Figure 2.11: Visualisation of ND gene set associations with the mitochondrial genome. A. Heatmap to show P-values associated with the median of four ND-related gene sets being more extreme than that of 10,000 random gene sets in 12 GTEx CNS regions. Raw P-values (below P<0.05) are represented by the colour scale, with the following asterisks overlaid to indicate which multiple test correction thresholds are passed: *0.05/12<P<0.05; **0.05/12*4<P<0.05/12; ***P<0.05/12*4. Grey squares indicate associations for which P>0.05. B. Visualisation of the results in part A, for the AOD target set only. The target gene set distribution is shown in blue and the distributions of 10,000 random size-matched gene sets are shown in green. Vertical dotted lines represent the medians of the target gene set (blue) and the central median of the 10,000 bootstrap sets (green). This figure was produced using the MitoNuclearCOEXPlorer tool[117].

nuclear correlations in 7/12 brain regions ($P<0.05$, Bonferroni-corrected for regions and gene sets), including the basal ganglia ($P<0.05$ for putamen, caudate and nucleus accumbens basal ganglia, Bonferroni-corrected for regions and gene sets) which are among the most disease-relevant tissues (Figure 2.12c). Similarly, genes associated with early onset dementia were found to have significant differences in mitochondrial-nuclear correlations in the majority of regions ($P<0.05$, Bonferroni-corrected for regions and gene sets).

Notably, in all cases, the ND-associated nuclear genes had more negative correlations with mtDNA gene expression than would be expected by chance. Interestingly, among the ND-implicated genes with the strongest mitochondrial-nuclear correlations was APP (in the top 1%, ranked 54/5898 of the negative mitochondrial-nuclear pairs), which encodes the precursor protein whose proteolysis generates amyloid beta ($A\beta$), the primary component of amyloid plaques. As well as this, highly significant mitochondrial-nuclear relationships were observed for some genes confidently associated with complex PD[65], such as PSAP (Figure 2.13). Interestingly, in PSAP knockout iPSC lines ROS production was seen to increase compared to controls[118]. As such, this analysis identifies high mitochondrial-PSAP association, lending support to this gene being important in core mitochondrial processes such as ROS-production.

Taken together, expression levels of genes causally implicated in a subset of NDs show stronger relationships with mtDNA gene expression than expected by chance. This analysis can be performed with a user-specified gene list using the accompanying tool available at the following URL: ainefairbrother-browne.shinyapps.io/MitoNuclearCOEXPlorer/.

2.3.6 Analysis of mitochondrial-nuclear correlations in Alzheimer's disease brain tissue

Finally, mitochondrial-nuclear correlations were analysed in post-mortem brain samples originating from individuals with AD and from matched neurological controls. The data were covariate corrected in the same way as for the GTEx data, but with the addition of Scaden-derived cell type proportions to account for disease-

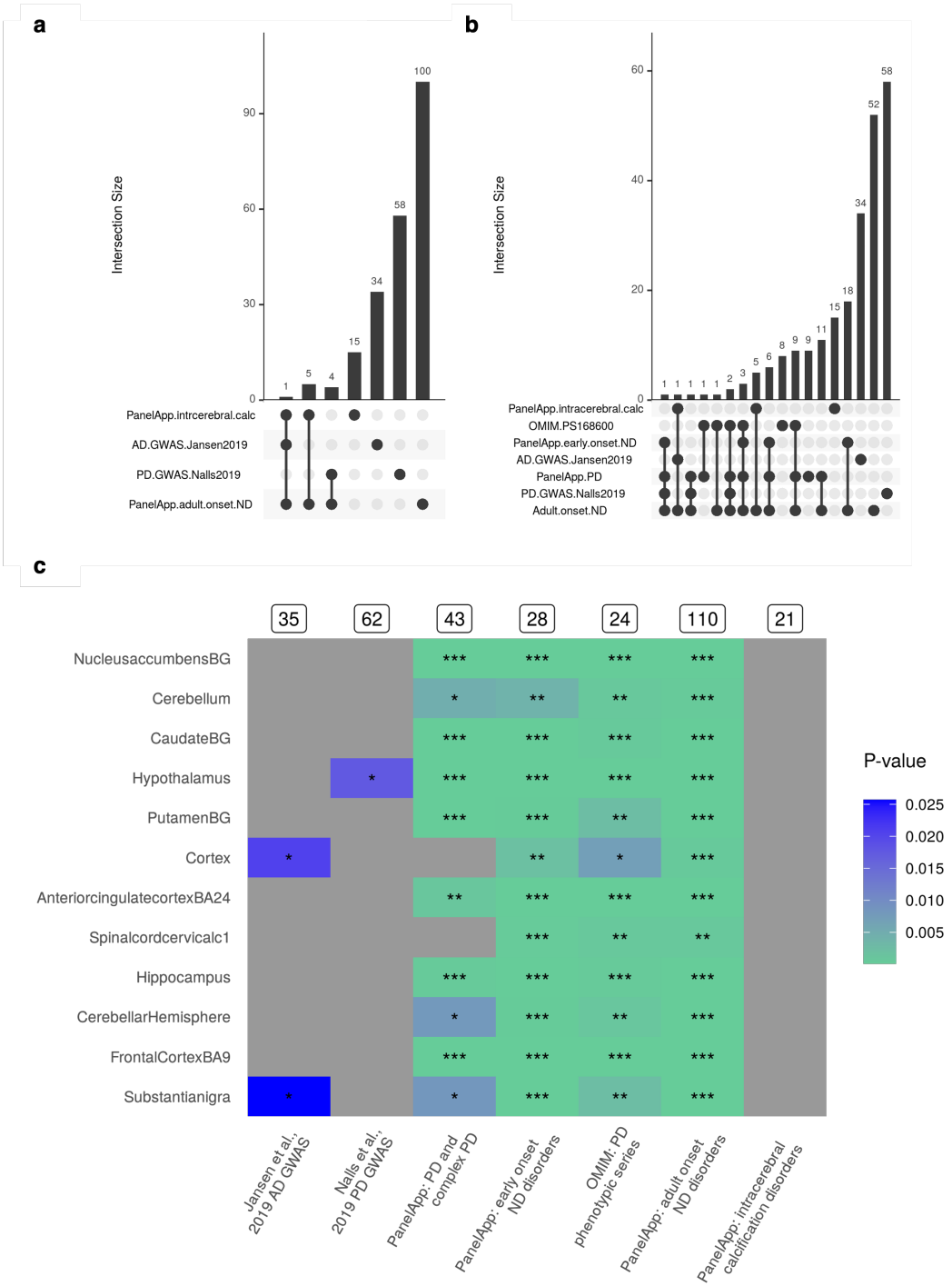

Figure 2.12: Determining whether ND-related gene sets have non-random associations with the mitochondrial genome. A. UpSet plot to show the overlap between the four gene sets included in the 'primary' disease gene analysis (see section 2.2.6 for gene set details). B. UpSet plot to show the overlap between gene sets included in the 'secondary' disease gene analysis (see section 2.2.6 for gene set details). C. Heatmap to show P-values associated with the median of 11 ND-related gene sets being more extreme than that of 10,000 random gene sets in 12 GTEx CNS regions (*0.05/12<P<0.05; **0.05/12*7<P<0.05/7; ***P<0.05/12*7).

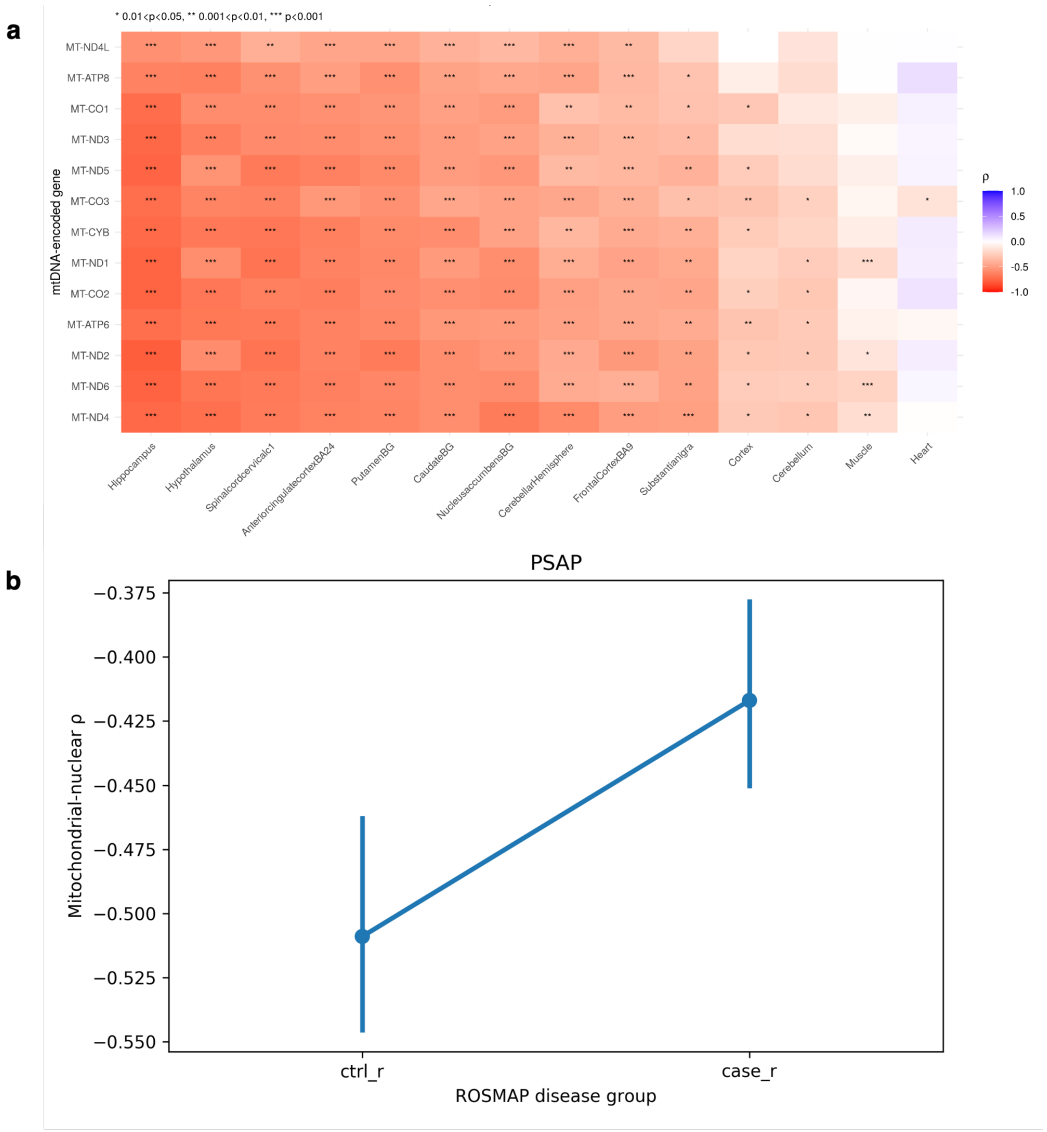

Figure 2.13: PSAP mitochondrial-nuclear correlations across GTEx tissues and in AD and control samples. A. MitoNuclearCOEXPlorer-generated visualisation of PSAP-mitochondria correlations across GTEx tissues. B. PSAP-mitochondria correlations derived from the ROSMAP dataset. Where Ctrl_r represents correlations (13 data points) in control samples, and Case_r represents correlations (13 data points) in control samples.

induced changes in cell type density. The difference in the correlation values between cases and controls for each mitochondrial-nuclear gene pair was calculated, resulting in case-control delta scores ($\Delta\rho$) (Figure 2.14a).

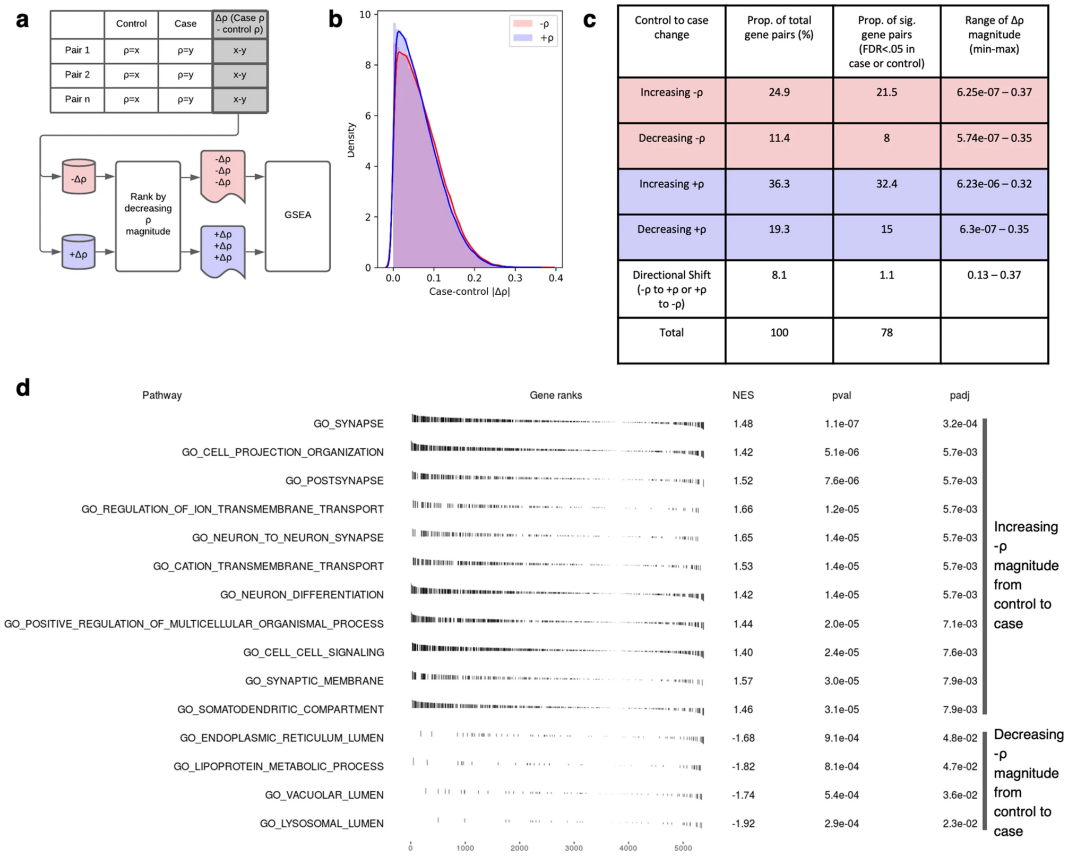

Figure 2.14: ROSMAP case-control analysis of ρ value differences ($\Delta\rho$). A. Schematic to show generation of the case-control $\Delta\rho$ values, splitting of the data into positive and negative $\Delta\rho$ values and subsequent ranking strategy applied prior to GSEA analysis. B Density plot to show the distribution of mitochondrial-nuclear case-control $\Delta\rho$ in ROSMAP frontal cortex data. The red curve represents negative $\Delta\rho$ and the blue curve represents positive $\Delta\rho$ values. C. Table to show the distinct groups of case-control $\Delta\rho$ values arising from the ROSMAP frontal cortex case-control data. D. fGSEA pathway enrichments passing P<0.05 (BH-corrected) for the negative correlation space, whereby gene pairs with -ρ have been ranked by their case-control $\Delta\rho$.

High levels of consistency between case and control mitochondrial-nuclear correlation values were observed, with 76% of pairs displaying a $\Delta\rho$ of <0.1 (Figure 2.14b). However, the presence of gene pairs displaying high delta scores was noted. In these pairs, coexpression had shifted in AD samples relative to controls (Figure 2.14b). Given that changes in cell type proportions had been corrected

for, these shifts likely represent disease-associated disruptions in mitochondrial-nuclear coexpression that have the potential to drive to AD pathogenesis. To understand whether nuclear genes involved in specific biological processes were represented amongst mitochondrial-nuclear gene pairs with high delta scores, Gene Set Enrichment Analysis (GSEA) was applied. First, gene pairs were split by their mitochondrial-nuclear correlation directionality, with the intuition that positive and negative correlations are representative of distinct transcriptional control mechanisms. 1.1% of significant shifts were observed among genes that switched directionality (Figure 2.14a) — these were excluded from the analysis. This yielded two gene sets (-$\Delta\rho$ and +$\Delta\rho$ scores), which were then ranked by their absolute $\Delta\rho$ score (Figure 2.14a).

In the negative correlation set, fGSEA detected 55 significant enrichments. The three most significant terms were synapse (P=3.5e-04, Bonferroni-Hochberg (BH) corrected), neuron to neuron synapse (P=4.6e-03, BH-corrected) and cell projection organisation (P=4.6e-03, BH-corrected), detected among gene pairs that display stronger relationships in case samples compared with controls. Three of the 55 enrichments (vacuolar lumen, and lysosomal lumen and lipoprotein metabolic process) were detected among gene pairs with negative mitochondrial-nuclear correlations that show weaker association in AD samples compared with controls. Within these sets, individual genes of specific interest for AD showed particularly large absolute $\Delta\rho$ scores. First, MTLN (rank 69/14,327 gene pairs with mean correlation taken across 13 mtDNA genes, ranked in the top 0.5% of $\Delta\rho$ values) encodes a protein product that is known to localise to the mitochondrial inner membrane, where it influences protein complex assembly and modulates respiratory efficiency, impacting on respiration rate, Ca2+ retention capacity and ROS[119,120], making it of particular interest in a disease context. Second, PSAP (max $\Delta\rho$=0.13, mean $\Delta\rho$=549/4653 ranked in the top 12% of decreasing -$\Delta\rho$ values) is a leading-edge gene for the lysosomal lumen enrichment and also displays highly significant mitochondrial-nuclear relationships across brain regions (Figure 2.13). This gene is of interest in the context of AD due to its known anti-inflammatory and neuro-

protective roles[121], as well as its identification as a biomarker of preclinical AD cases, enabling discrimination from control samples[122]. No enrichments reaching BH significance were detected in the positive correlation list.

2.3.7 The MitoNuclearCOEXPlorer tool

Alongside the publication containing this work[117], a tool was released enabling analyses performed in the publication to be performed with a user-specified gene list (Figure 2.15). The tool allows for: (i) single gene querying of the correlation data, (ii) assessment of association with mtDNA gene expression with a gene list (as in Figure 2.11), (iii) generation of publication-quality visualisations for all analyses and download of processed data. The MitoNuclearCOEXPlorer tool can be found at the following URL: ainefairbrotherbrowne.shinyapps.io/MitoNuclearCOEXPlorer. The accompanying source code can be found at github.com/ainefairbrother/MitoNuclearCOEXPlorer.

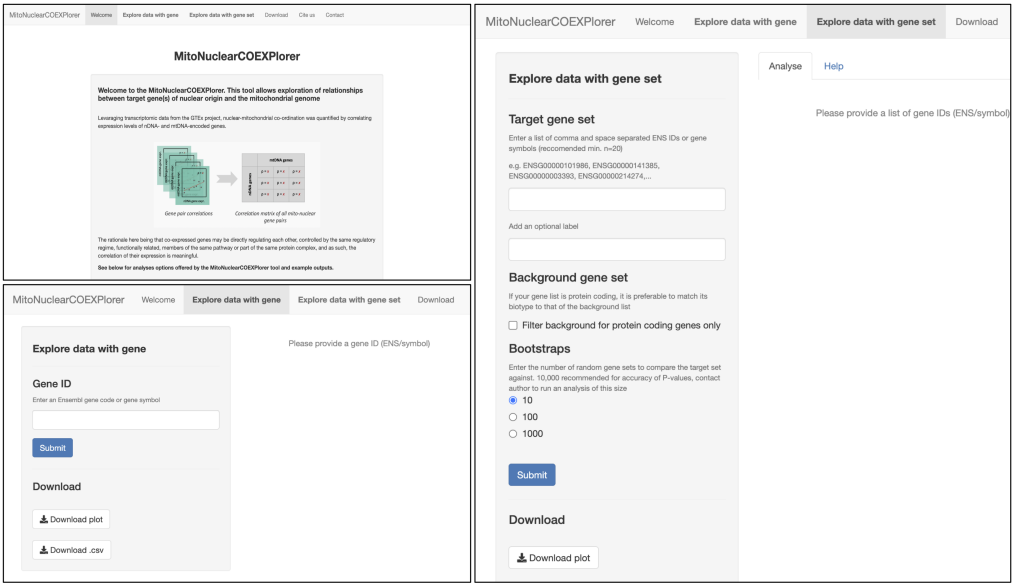

Figure 2.15: Screen captures demonstrating the MitoNuclearCOEXPlorer web application. The top left pane is a screen capture of the home page, which gives aims, methods and usage examples for each analysis available within the tool. The bottom left pane shows the interface for analysis (i) described above. The right-hand pane shows the interface for analysis (ii) described above.

2.4 Discussion

This work aims to understand mitochondrial-nuclear coordination patterns across the human CNS. CNS regional cell type composition was found to contribute to regional variation in coexpression, reflecting functional specialisation, specifically at synapses. Using an independent frontal cortex dataset, high replicability of mitochondrial-nuclear correlation distributions and cell type-specific correlation profiles was demonstrated. Nuclear genes causally implicated in PD and AD were found to have much stronger relationships with the mitochondrial genome than expected by chance. Finally, mitochondrial-nuclear relationships were found to be highly perturbed in AD cases, particularly those involving synaptic and lysosomal genes.

A key finding of this study was the identification of cell type as a contributor to the distinct patterns of mitochondrial-nuclear coexpression across CNS regions. Neuronal markers were enriched in negative mitochondrial-nuclear correlations, in contrast to glial (astrocytic and microglial) markers which were enriched in positive correlations. Additionally, a reduction of cross-CNS variation in mitochondrial-nuclear correlations was observed when correcting for cell type proportions, indicating that depletion of cell type -specific signals reduces the regional specificity of mitochondrial-nuclear relationships. Notably, correction for cell type significantly, but not entirely, depleted cross-regional variation, indicating that although cell type is a significant contributor, there are additional drivers of cross-regional variation in mitochondrial-nuclear relationships.

The finding that cell type significantly contributes to regional variation in mitochondrial-nuclear association could be explained by cell type-specific mitochondrial specialisation. These analyses assay a proxy for the nuclear association with ATP synthesis, and so capture a single aspect of mitochondrial function. In fact, mitochondria have many other important roles in cells, such as calcium buffering, which may vary across different cell types. As such, the division of mitochondrial-nuclear correlation directionality between cell types could be the result of divergent functionality among the mitochondria of these cell types. This is a

view supported by proteomic cell type -specific profiling of brain mitochondria. Recent work has revealed notable molecular and functional diversity of mitochondria across cell types, with astrocytic mitochondria found to perform the core cellular functions of long-chain fatty acid metabolism and calcium buffering with greater efficiency than mitochondria in neural cell types[41]. Another linked explanation for cell type -specific correlation directionality is that it is driven by core differences in energy management strategies between cell types. In energetically demanding cell types such as neurons, anti-correlation could reflect the need for tighter OXPHOS regulation to protect against excessive ROS production, with post-transcriptional processes perhaps being used to manage local, flexible regulation of energy supply. Interestingly, oligodendrocytes were the exception among the glial cell types, displaying neuron-like enrichment in negative mitochondrial-nuclear correlations. In this context, it is worth noting that while oligodendrocyte metabolism is poorly understood, their central role in myelin sheath production is highly energy intensive, mirroring the high energy requirements of neurons[123,124].

The synapse is the site of greatest energy expenditure in the neuron[125]. To match energy supply and demand, the mitochondria in synaptic compartments display structural, biochemical and spatial plasticity[125]. To achieve this necessitates equally flexible maintenance of the mitochondrial proteome, the exact mechanisms of which are not known[125]. The analysis set out in this chapter reveals variable mitochondrial-nuclear relationships being highly significantly enriched for synaptic marker genes, meaning that nuclear-encoded synaptic gene expression and mitochondrial-encoded gene expression are differentially associated across the CNS. The possibility was considered that this is simply tagging variability in regional mtDNA expression; however, residual TPM values for the 13 mtDNA genes demonstrate consistent cross-CNS expression (Figure 2.8a), suggesting that this is not a core driver of the regional specificity of mitochondrial-nuclear correlation profiles. It could be that this represents mitochondrial plasticity, where neuronal subtype-specific variation in nuclear and mitochondrial expression is being captured. It is known that neuronal subtypes are energetically specialised, and that

CNS tissues have differential neuronal subtype compositions[126,127,128]. Sub-cell type -specific expression modulation as a mechanism to manage local energy requirements at synapses is supported by work finding that heterogenous energy requirements across CNS regions and cell types may necessitate bespoke mitochondrial proteomes[125]. Further to this, molecular evidence shows that several nuclear-encoded mitochondrial genes involved in processes key to mitochondrial plasticity (mitochondrial transcription, fission and trafficking) have been found to exhibit distinct patterns of expression in neuronal subtypes[129]. Recent work using engineered MitoTag mice coupled with an isolation approach to profile tagged mitochondria from defined cell types has demonstrated profound cell type-specific mitochondrial biology serving homoeostatic needs to preserve essential functions in cells[41]. And yet, without directional information and cell type or sub-cell type -specific data, it is difficult to make a firm assertion as to whether the underlying mechanism is anterograde modulation of the mitochondrial genome from the nucleus, or retrograde modulation of the nuclear genome by the mitochondria, or perhaps a feedback loop involving both.

Uniquely to the field of mitochondrial-nuclear cross-talk, this chapter looks at its genome-wide relevance with respect to a range of NDs. Testing the association of ND-implicated genes with the mitochondrial genome demonstrated significant non-random correlations between mtDNA gene expression and ND-implicated nuclear genes. While genes implicated in PD and AD through GWAS analyses showed nominally significant associations with the mitochondrial genome, it should be noted that as well as potential inaccuracies in variant-gene assignments within these sets, the effect sizes of the variants nominating the genes is likely to be small, weakening the analysis. Interestingly, this view is supported by high confidence enrichments of mitochondrial-nuclear association in nuclear gene sets associated with Mendelian forms of the same diseases. Mendelian AD and PD genes displayed highly significant shifts from random, all of which were towards higher negative correlation magnitudes, and highlighted particularly strong correlations among important ND genes. In fact, APP, the first gene to be causally implicated in AD,

ranked in the top 1% of all pairs with negative associations.

Given these findings, the hypothesis was that analysing changes in mitochondrial-nuclear correlations in the context of AD would provide important disease insights. To look at this, the AD case-control ROSMAP dataset was leveraged. After correcting for cell type proportion, an enrichment of synaptic terms among nuclear genes which were negatively correlated with mtDNA gene expression was observed. These genes had stronger relationships in the context of AD than in control samples (i.e. high case-control correlation difference, $\Delta\rho$, gene pairs). Given the close relationship between synapses and mitochondria, with multiple lines of evidence pointing not only to synaptic function being dependent on mitochondria, but to mitochondrial regulation of synaptic plasticity, the tightening coexpression here could represent a drive to recover energetic homoeostasis at damaged synapses and increase their efficiency[130,131,132]. In support of this, mitochondrial efficiency enhancing gene *MTLN31* was in the top 1% of increasing negative associations. In particular the *MTLN-MTCYB* gene pair displayed a striking $\Delta\rho$. In control samples the pair had a non-significant correlation (ρ=-0.008, P=0.93), but shifted to a highly significant association with a considerably higher negative magnitude in case samples (ρ=-0.27, P=3.01e-05).

Interestingly, enrichment of lysosome-related terms (lysosomal lumen, vacuolar lumen) was observed in negatively correlated gene pairs that weaken in case samples relative to controls (Figure 2.14d). Lysosomes are essential for the removal of dysfunctional mitochondria as well as other organelles and proteins, and there is growing evidence to suggest that lysosomal dysfunction contributes to the pathogenesis of AD, as well as PD[133,134,135]. Perhaps decoupling of nuclear genes in these pathways from mtDNA gene expression represents a reduction in the efficacy of dysfunctional mitochondria clearance, thus augmenting the pathology.

While this study yielded a number of robust findings, it is important to acknowledge its limitations. One potential concern is the confounding effect of ancestry composition in the datasets used. Although both GTEx and ROSMAP are largely comprised of individuals of European ancestry, some stratification does ex-

ist, wherein both datasets contain non-trivial numbers of African American donors and small proportions of donors with other ancestries[136,100]. Population stratification has the potential to introduce systematic biases into transcriptomic studies, and is particularly important with respect to studies of the mitochondrial transcriptome. This is due to the fact that human populations can contain distinct mitochondrial haplogroups whose mtDNA variants can impact levels of mtDNA-encoded transcripts. To control for population substructure whilst retaining the full power of the dataset, a common approach is to utilise genome sequencing data to calculate genetic principle components (gPCs). These are eigenvectors that each explain a proportion of inter-individual genetic variation in the dataset and gPCs can be corrected for in the same manner as other covariates. However, this approach relies on the presence of paired DNA and RNA sequencing for all participants, which is unfortunately not the case for the cohorts utilised in this study.

A second potential limitation that is important to consider is the normalisation strategy utilised. In this study, read counts were converted to TPMs to control for gene length and library size. TPMs were then median normalised to improve inter-sample comparison and log10 transformed to remove skewness. Although this is a standard approach utilised throughout the transcriptomic literature and in studies similar to the present one[46,96], future studies may consider a more tailored approach, as normalisation methods are known to have a large impact on downstream interpretation of results. For example, an assumption of TPM normalisation is that library composition is consistent across samples[137]. However, the percentage of mtDNA-encoded reads is known to be variable between samples, even in the same tissue. As such, in some samples, mitochondrial reads may constitute a large proportion of the library, and in others they may constitute a smaller proportion[137]. Thus, in the former, nDNA-encoded transcripts would constitute a small proportion of the library and in the latter a larger proportion. In this way, correlations derived from mtDNA and nDNA -encoded transcript counts that are TPM normalised have the potential to be confounded by library composition[137]. Given this, in-depth assessment and subsequent use of normalisation methods that consider library com-

position among other factors may improve the accuracy of downstream correlation coefficients in future studies of mitochondrial-nuclear transcriptomic coordination.

This work surveys the mitochondrial-nuclear relationship broadly, across CNS tissues and clinical states, providing further evidence for the role of mitochondrial-nuclear co-ordination in ND. This chapter begins to examine disruption of the nuclear-mitochondrial relationship in AD, the most common ND, for which a substantial post-mortem brain RNA-Seq dataset is available (ROSMAP). Leveraging this, this work identifies, in the disease-relevant tissue, pathways and processes that may be relevant to AD aetiology and could represent routes to dysfunction of that are of particular importance for the development of disease modifying treatments. However, as is the consensus in the literature, and is supported by these results, it is PD, the second-most common ND, for which mitochondrial dysfunction is thought to be a more central component. As such, understanding the involvement of the mitochondrial-nuclear relationship may be particularly pertinent. Thus, in the chapters that follow, the focus will shift to surveying the mitochondrial-nuclear relationship in PD, utilising the Accelerating Medicines Partnership for Parkinson's (AMP-PD) project dataset[138]. This project harmonises clinical and sequencing data across individual PD cohort studies, resulting in the availability of high quality, deeply sequenced, whole blood RNA-Seq data across many clinically profiled individuals and at multiple time-points for PD cases and healthy controls. As such, this is a highly valuable dataset within which to interrogate the mitochondrial-nuclear relationship in a pre-mortem tissue. The statistical power available through this dataset permits expanding beyond looking at the mitochondrial-nuclear relationship through the lens of coexpression, allows scrutiny of mitochondrial transcriptional control across several forms of PD and makes it possible to study nuclear genetic effects on mitochondrial expression, adding a directionality to the study of this relationship.

Chapter 3

Understanding mitochondrial transcription in monogenic and sporadic Parkinson's Disease

3.1 Introduction

Monogenic PD is caused by variation in a single gene and is characterised by marked familial inheritance patterns. In contrast, sporadic (or idiopathic) PD is associated with variants that collectively increase an individual's risk, in addition to environmental and environment-by-gene interaction contributions. Interestingly, the two forms are generally clinically indistinguishable[139] and genome wide association studies (GWAS) focused on characterising the genetic component of sporadic PD have uncovered genes associated with monogenic PD[140]. This suggests the existence of shared pathogenic mechanisms, pointing to the idea that a deeper understanding of the aetiology of monogenic forms could yield insights into those of the sporadic form.

Despite ongoing research efforts, however, the underlying causes of sporadic PD remain largely unknown. Equally, the mechanisms linking the variant to the phenotype remain unclear in monogenic forms. Notably, there are currently no effective treatments that can prevent or intervene in PD progression, making understanding disease mechanisms and identifying druggable targets crucial. Increas-

ing evidence points to mitochondrial dysfunction as playing an integral role in the pathogenesis of both monogenic and sporadic PD, including known roles for monogenic PD genes in mitochondrial processes (*PINK1*, *PARK2*, *PARK7*, *CHCHD2*, and *VPS13C* have roles in mitochondrial quality control)[141]. However, considerable work remains to fully define the mechanisms underlying the pathogenic influence of mitochondrial processes both in monogenic and sporadic forms[142]. Three essential aspects remain unclear, firstly, the degree to which mitochondrial dysfunction contributes to PD pathogenesis and in which patients. Secondly, whether it does so early or late in the disease, and thirdly, whether mitochondrial dysfunction is an important pathophysiological process in all patients or only a subset. Currently, even symptomatic treatments have high variability across patients both in terms of efficacy and off-target effects[143]. It is thought that the existence of clinical subtypes arising from divergent pathogenesis may be a factor here[144].

Consequently, a more precise understanding of the role of mitochondrial dysfunction has implications for the development of biomarkers and the identification of novel drug targets, two key areas of interest in PD research at present[61]. Novel biomarkers indicative of molecular or progression status will enable better counselling, care and treatment of patients[61]. Their development is also important to clinical trial design by informing participant inclusion and enabling the selection of individuals most likely to respond to the putative disease modifying drug[144]. In turn, this has the potential to permit a more accurate assessment of drug efficacy in clinical trials and to generate more favourable outcomes for patients. As such, better understanding of not only the contribution, but also the timing of mitochondrial dysfunction bridges an important knowledge gap and has practical applications.

The approach taken in this chapter was to compare transcriptomic profiles across participant groups defined by PD form (sporadic or monogenic) and symptomatic status (asymptomatic or symptomatic). The transcriptomic profiles consisted of mtDNA and nDNA -encoded OXPHOS gene expression. OXPHOS is a core mitochondrial process which has previously been linked to PD, as such this was seen as a useful way to assess the function of an important mitochondrial process in

the context of PD.

The AMP-PD dataset[138] is an ideal resource in which to explore the questions posed in this chapter for several reasons. Firstly, it is comprised of whole blood -derived 'omics data, providing a number of advantages over often favoured post-mortem brain tissue data. Blood samples are relatively noninvasive, making it easier to generate a resource that contains large sample numbers. Secondly, and coupled with this, the samples are obtained from living individuals, avoiding many of the issues associated with post-mortem tissue such heterogeneity of sample quality and RNA degradation. Biologically speaking, the body-first model of PD implicates body tissues such as the gut and immune cells in PD pathogenesis[145]. As such, it is expected that blood, through its interaction with these affected cells, would capture signals of PD. Thirdly, the participants in the AMP-PD dataset fall into four clinically defined groups, which is useful for testing a range of hypotheses. The main groups of interest in this study are healthy controls, sporadic PD, symptomatic monogenic PD (*LRRK2*, *GBA*, and *SNCA*), and asymptomatic monogenic PD (*LRRK2*, *GBA*, and *SNCA*). This structure is particularly useful for three main reasons; (i) the presence of both monogenic and sporadic forms of PD provides an opportunity for the comparison of monogenics forms (with known genetic causality and inferred molecular mechanisms) with the sporadic form whose pathogenesis remains unclear, (ii) the availability of multiple 'unified' cohorts (processed using the same pipeline) within one dataset allows for intra-dataset discovery and replication and, (iii) the asymptomatic and symptomatic monogenic PD participants provides a paradigm within which to explore questions around disease-stage and the timing of aspects of pathogenesis.

To answer the questions raised in this context, two main aims were established; (i) to study the role of mitochondrial transcription in PD pathogenesis and, (ii) to determine whether mitochondrial transcriptomic profiles change in disease, whether they are able to distinguish between sporadic PD cases, monogenic PD cases and healthy controls. Then, if changes are observed, to understand at what point in the disease trajectory these arise. Overall, this chapter aimed to capitalise on the unique

opportunities provided by the AMP-PD dataset to better understand the molecular mechanisms underlying PD. By exploring the transcriptional profiles of mtDNA and nDNA -encoded genes across monogenic and sporadic PD patients, it is hoped that new insights into the mitochondrial component of PD pathogenesis can guide the development of novel treatments for the disease.

3.2 Methods

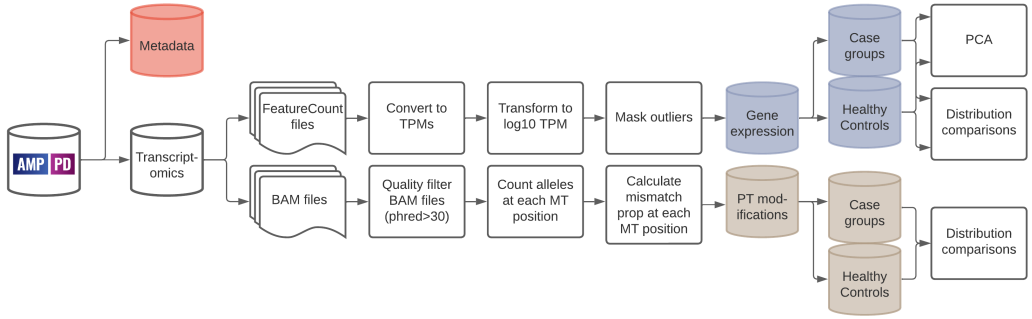

Figure 3.1: Workflow diagram to summarise the data and methods utilised in chapter 2.

3.2.1 The AMP-PD dataset

The AMP-PD data[138] utilised throughout this chapter and the next were accessed through a request to the Broad Institute for 'Clinical' level access, which permits access to the genomic, transcriptomic and clinical data. Data were either directly analysed on the Broad Institute's cloud platform, Terra (https://app.terra.bio/), or downloaded from the AMP-PD Google Cloud storage bucket and analysed locally. The data release used was the 2019 V1 release, the latest available when analysis commenced.

The Accelerating Medicines Partnership (AMP) program is a public-private partnership between the National Institutes of Health (NIH) and other industrial partners, aiming to identify promising druggable targets. A major disease of focus is PD, for which 'omics data for thousands of patients and healthy controls are collated in the AMP-PD project, which is leveraged in this chapter and the next. The specific goals of the AMP-PD project include molecular characterisation of the disease and longitudinal profiling of PD patients, seeking to identify biomarkers,

dissect new targets, identify disease subtypes and predict disease progression. To this end, the project has collected and harmonised the data of several separate PD studies. Starting at the level of the blood sample, consistent laboratory and data processing pipelines were applied to collect Whole Genome Sequencing (WGS) data and RNA-Seq data. Clinical information was also collected in a consistent manner.

In this work two cohorts from the 1.0 release of the AMP-PD dataset were considered - Parkinson's Progression Markers Initiative (PPMI)[146] and Parkinson's Disease Biomarkers Program (PDBP)[147]. For these cohorts, clinical, RNA-Seq and WGS data was available for a large number of participants across multiple timepoints, meaning that these cohorts, of all AMP-PD cohorts, had the greatest power for downstream analyses (total participant numbers: PPMI=1478, PDBP=1467). It was important to consider the composition of these datasets to inform experimental design and as such, the original aims of the PPMI and PDBP studies, structure and inclusion criteria are detailed in the sections that follow - 3.2.1.1 and 3.2.1.2.

3.2.1.1 General outline of the PDBP cohort

The PDBP cohort[147] was originally curated by National Institute of Neurological Disorders and Stroke (NINDS), and was focused on discovery of new biomarkers for PD (https://pdbp.ninds.nih.gov/). The general inclusion criteria for this study required that the participant must be a male or female ages 21 years or older. General participant exclusion criteria included the following: on anti-coagulants, a history of neuroleptic use, a history of schizophrenia or the use of investigational drugs within 60 days prior to the baseline visit (first clinical visit of the study). Specifically for individuals to be considered as controls, they must not have a 'current or clinically significant neurological disorder' as per the assessing clinician and must not have a family history of ND in a first or second degree blood relative. For individuals to be included as having PD they must have had a primary diagnosis of PD at the time of recruitment (table A.2 in Appendix A).

3.2.1.2 General outline of the PPMI cohort

The PPMI cohort[146] data is derived from a Micheal J. Fox foundation and industrial partner -sponsored project launched in 2010 (https://www.ppmi-info.org/). This longitudinal observational study has collected genetic and clinical data from participants over a time period spanning between 5 and 13 years, utilising 33 clinical sites around the world. The primary focus is to identify biomarkers of PD progression for use in clinical trials and to develop disease-modifying therapies. The PPMI dataset has a more complex structure than the PDBP dataset, containing seven different categories of participant: PD, healthy control, 'SWEDD' (scans without evidence for dopaminergic deficit), 'Prodromal', genetic cohort affected, genetic cohort unaffected and 'Genetic Registry'. For each category, specific inclusion and exclusion criteria apply (a detailed breakdown of the criteria for each participant category can be found in table A.2 in Appendix A). To briefly summarise this information, the PD category contains sporadic, early stage, symptomatic PD cases that are not on any PD medication. The healthy control category contains neurologically healthy individuals with no instances of PD in a first degree relative. The 'SWEDD' category is a sub-set of the PD category, requiring participants to have no evidence of dopamine transporter deficit on a SPECT scan. The 'Prodromal' category is also a sub-set of the PD category, but requires the participants to be older than 60, and to have a rapid eye movement (REM) sleep disorder. The genetic cohort affected category includes mid- to late- stage symptomatic participants aged 18 years or older who have a variant in *LRRK2*, *GBA* or *SNCA*. Contrastingly, the genetic cohort unaffected category contains individuals 45 years or older who have a *LRRK2* or *GBA* variant, or 30 or older with an *SNCA* variant. Importantly, genetic cohort unaffected individuals have no clinical diagnosis of PD. Finally, the 'Genetic Registry' category contains participants that have, or have a first degree relative with, a *LRRK2*, *GBA* or *SNCA* variant — AMP-PD listed no exclusion criteria for this category.

3.2.1.3 Participant composition in the PPMI and PDBP cohorts

The AMP-PD project has performed cross-cohort harmonisation across all included studies, yet there are essential differences between PDBP and PPMI participants.

These were important to note when considering their joint usage in downstream analyses. The most basic difference is that PDBP consists only of PD, healthy control and 'Disease Control' participants, whereas PPMI does not contain 'Disease Control' and additionally contains Genetic, 'Prodromal' and 'SWEDD' participants (Fig. 3.2c). In terms of tracking participants over time, PDBP records data at baseline (M0) through M6, M12, M18 and M24, whereas PPMI records data at M0, M6, M12, M24 and M36. PPMI tracks a maximum of three years for each participant which is higher than PDBP which tracks a maximum of two years, but at a marginally higher resolution due to data collection at visit M18 (Fig. 3.2c). It is important to note, however, that the timepoints do not carry the same meaning across cohorts. This is best illustrated by the medication data, wherein PD cases are recruited to PPMI within 6 months of a diagnosis. In the PDBP cohort however, all participants are already on medication, reflecting a slightly later disease stage. Hence, at 'M0', PDBP and PPMI participants will be at differing disease stages.

Age and sex are known to be important with regard to PD risk and progression, so it was salient to note the distributions of these in the two cohorts. In the PD category, the median age of the PDBP cohort was 62 (range=34-87) whereas for PPMI it was 58.5 (range=34-85) (Fig. 3.2a). For the healthy control participants the median age of the PDBP cohort was 57 (range=23-86), and the median age of the PPMI cohort was 59.5 (range=31-84) (Fig. 3.2a). As such, the age profiles of the two cohorts are well matched for PD and healthy control participants. The sex ratios are similar across cohorts for PD participants, where these are 37% female for PDBP, and 35% female for PPMI (Fig. 3.2b). For the healthy control participants, 57% of participants are females for PDBP, whilst for PPMI this figure is 35% (Fig. 3.2b). Thus, the sex ratios are well matched within PPMI across cases and controls, but not as well matched within PDBP. Across the cohorts, sex ratios are well matched for cases, but not as well matched for controls.

Another important aspect of the data is the distribution of medication use, as the medication usage itself may influence the underlying biological and transcriptional state of an individual with disease. Four categories of medication usage are

recorded in the PPMI and PDBP datasets. These are: 'Dopamine agonist', 'Levodopa', 'Other PD medication' and 'None' (Fig. 3.2c). In some cases, participants are assigned 'NA' values, a label which is intuitive for control participants, but is seemingly ambiguous for cases. In the PDBP cohort, all PD participants are on medication, with the prevailing medication being Levodopa. A small proportion of participants are on dopamine agonists, and a smaller proportion still are on 'Other PD medication'. The ratio of participants on each medication is approximately consistent across visits in the PDBP cohort (Fig. 3.2c). In the PPMI cohort, medication usage is variable across participant categories and across visits (Fig. 3.2c). Across visits, PD case participants shift from predominantly no medication in M6 to the majority taking Levodopa by M36. Additionally, a non-trivial proportion of PPMI participants are on dopamine agonists by M12.

All in all, the PDBP and PPMI cohorts display differences in their essential compositions, including in the participant categories, sex ratios, timecourse structures and medication usage. There are, however, comparable subsets of the cohorts. For example, the healthy control category is relatively consistent between PPMI and PDBP at both the M12 and M24 timepoints. As well as this, participant medication profiles are similar between PDBP M0 and PPMI M36.

3.2.1.4 PPMI participant selection strategy

Due to the complexity of participant disease classifications in the PPMI dataset, the data was systematically filtered to select appropriate samples for downstream analyses. In this chapter, the analyses focus on PD cases and healthy controls at a single timepoint. Month 12 (M12) was selected as it was the most complete timepoint (Fig. 3.3d). The largest number of participants had data at M12, including those in the genetic cohort, PD and healthy control categories. To gain power across participant categories, and in particular the PD category, the M12 participants were combined with 'M24 orphan' participants. These are participants for which data was collected at M24 but not at M12. The dataset generated using M12 plus M24 orphans is shown in figure 3.3b, demonstrating a total addition of 143 participants across all participant categories. The assumption here is that these timepoints are

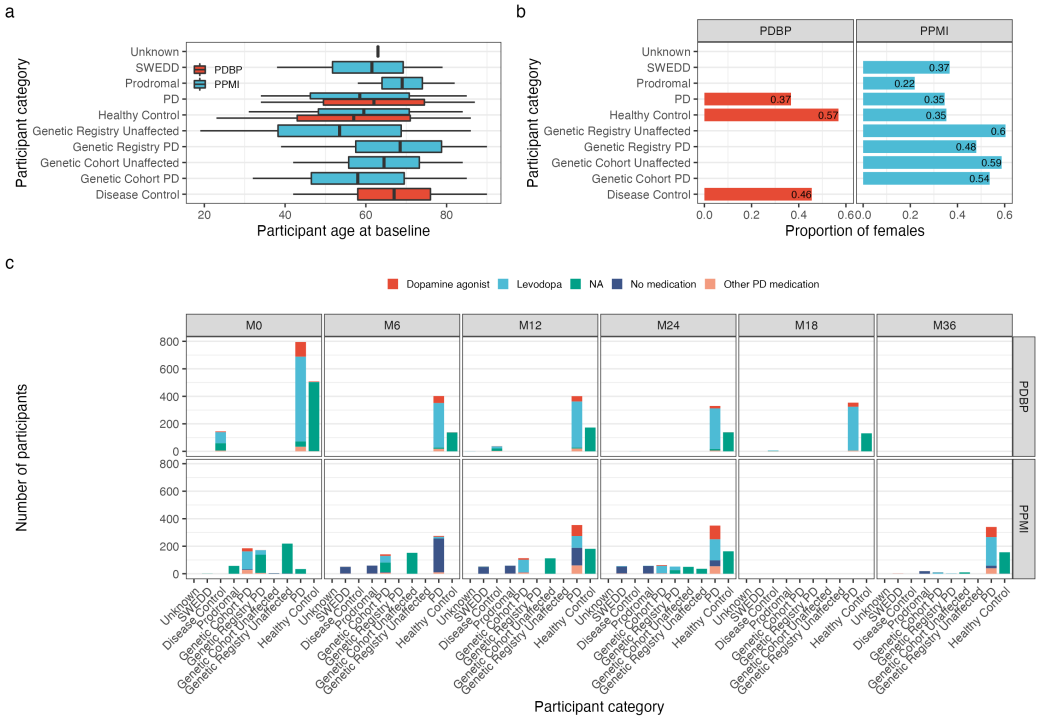

Figure 3.2: Comparison of participant composition of the AMP-PD -derived PPMI and PDBP cohorts. A. Distributions of participant ages, stratified by participant category, and coloured by cohort, where PDBP is shown in red and PPMI in blue. The 'SWEDD' category is only present in the PPMI dataset and contains participants that have 'scans without evidence of dopaminergic deficit'. The 'PD' category contains participants diagnosed with sporadic PD. The 'Genetic Cohort' category (unaffected, PD) participants have WGS-confirmed genetic PD mutations in *LRRK2*, *SNCA* or *GBA*. The 'Genetic Registry' category (unaffected, PD) participants have s first or second degree relative with a WGS-confirmed mutation. B. Proportions of female participants in the PPMI and PDBP datasets, stratified by participant category. C. Sample counts, coloured by medication status, by participant category (on the x-axis, as described in A), timepoint (where 'M' indicates month and 'M0' indicates the first sample taken from an individual) and cohort (PDBP and PPMI).

only 1 year apart and so the relatively slow progression of PD means they can be considered equivalent. In addition, in the AMP-PD data, timepoints (visits) represent disease progression of the individual, but this may not be generalisable across participants. For example, the Genetic Cohort criteria stipulate a Hoehn and Yahr stage of less than IV at baseline. Hypothetically, this could mean that at M0, one participant is at stage III, where another is at stage I. Both participants are thus assigned 'M0' at different points in their respective disease trajectories. In light of this, combining across M12 and M24 was seen as a reasonable strategy for increasing sample number, and by extension, statistical power, for downstream analyses.

Participant filtering was also carried out to remove those with conflicting labels as follows: participants labelled 'Case' in the case_control_other_latest variable, and healthy control in the study_arm (participant category) variable, participants labelled 'Control' in the case_control_other_latest variable, and PD in the study_arm variable or had a known PD variant. Finally, participants labelled PD in the study_arm and had a known variant were filtered out to ensure the PD category contains purely sporadic cases. This sample selection strategy resulted in a final dataset of 614 unique participants, including 108 genetic cohort PD (*GBA*=16, *LRRK2*=80, *SNCA*=12), 170 healthy control and 336 sporadic PD participants.

3.2.1.5 PDBP participant selection strategy

Full details of the PDBP participant selection strategy are supplied in section 4.2.1.1. However, to summarise, of 1467 total PDBP participants available, after filtering, 1291 remained, including 507 in the "Healthy Control" group and 784 in the "PD" (sporadic PD) group.

3.2.2 RNA-Seq data processing

The analyses in this chapter and the next leveraged two mitochondrial transcriptional phenotypes, both derived from AMP-PD RNA-Seq data: expression levels of 15 mtDNA-encoded genes and post-transcriptional (PT) modification rates at 21 sites. The workflow diagram in figure 3.1 gives a broad overview of the steps taken. Additionally, the detailed participant and time-point selection rationale is outlined

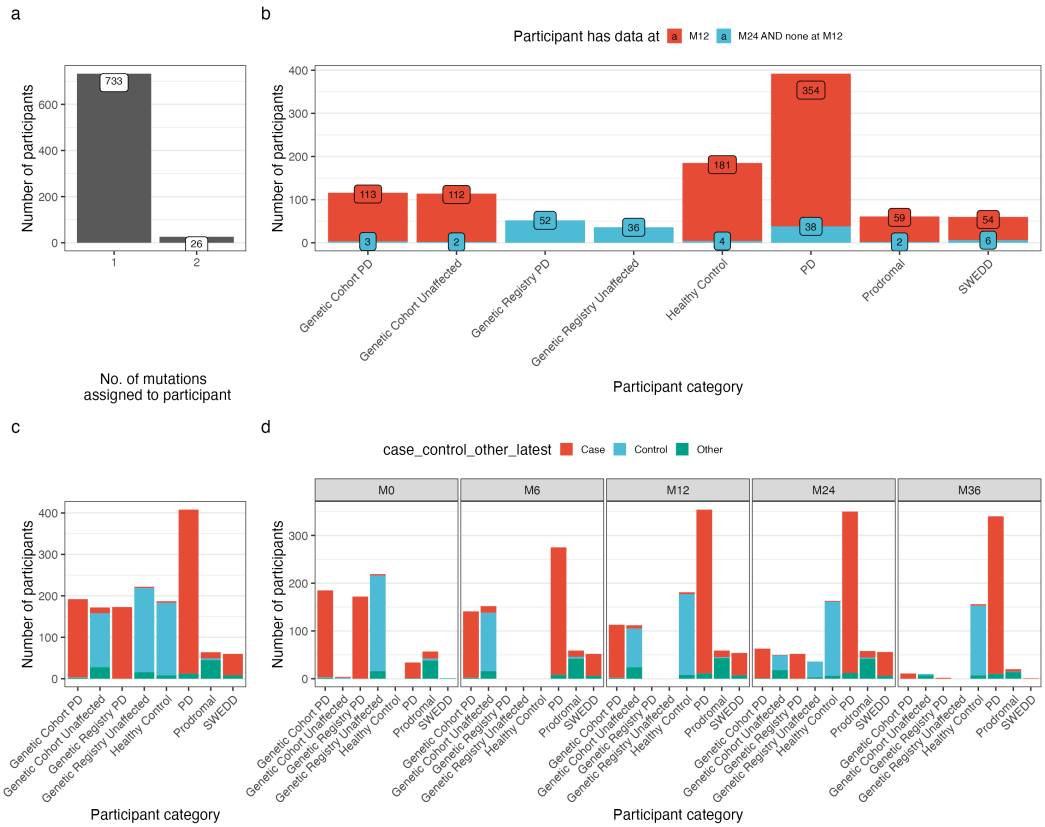

Figure 3.3: Participant composition of the AMP-PD -derived PPMI dataset. A. Barplot to show the of participants assigned >1 PD variant. B. Barplot to show, for each participant category, the number of participants with data at timepoint M12 (red), and the numbers gained by adding those with data at M24 but not M12 (blue). Exact numbers shown on labels. C. Barplot to show participant numbers in each category, coloured by case/control/other status. D. Same as in C, but stratified by timepoint.

in section 3.2.1, resulting in the selection of PPMI-M12 and PDBP-M0 participants to take forward in downstream analyses. The particulars of the data generation, procurement, filtration, correction and analysis methods are outlined in the following sections.

3.2.2.1 Gene expression data

The AMP-PD project provides raw data as well as data processed using various standard bioinformatic tools. AMP-PD RNA-Seq data was sequenced on the Illumina NovaSeq 6000 platform, generating 100M paired reads per sample with 150nt read lengths. Alignment was performed using STAR and the GRCh38 hu-

man genome build. STAR was run using the following options, as per the AMP-PD trancriptomics workflow[138]:

```
STAR --genomeDir STARREF --runMode alignReads
--twopassMode Basic\
--outFileNamePrefix SAMPLEID --readFilesCommand zcat\
--readFilesIn FASTQL1 FASTQL2
--outSAMtype BAM SortedByCoordinate\
--outFilterType BySJout --outFilterMultimapNmax 20\
--outFilterMismatchNmax 999
--outFilterMismatchNoverLmax 0.1\
--alignIntronMax 1000000 --alignMatesGapMax 1000000\
--alignSJoverhangMin 8 --alignSJDBoverhangMin 1\
--chimOutType WithinBAM --chimSegmentMin 15\
--chimJunctionOverhangMin 15 --runThreadN 16\
--outSAMstrandField intronMotif
--outSAMunmapped Within\
--outSAMattrRGline RGTAGLIST
```

The available processed data included featureCounts[148] files generated by counting aligned reads from the STAR-generated BAM files. It is important to note that featureCounts by default only counts reads that are properly paired and uniquely mapped, minimising the chance of signal arising from nuclear-encoded fragments of mitochondrial DNA (NUMTs) and mtDNA cross-mapping. featureCounts (v1.6.2) was run using the following options, as per the AMP-PD trancriptomics workflow[138]:

```
--T 2 -p -t exon -g gene_id
--a gencode.v19.annotation.patched_contigs.gtf
--s 2
```

Read quantification files generated by the featureCounts tool were downloaded from the AMP-PD Google Cloud bucket. All processing and analyses concerning

the count data were carried out using R (v4) on a local server thereafter. Transcripts per million (TPM) values were calculated from the read counts to normalise for sequencing depth and gene length. TPMs were then log10 transformed ($log10(TPM+1)$) to allow for more accurate comparison of transcript abundance between samples. log10 TPMs were split into separate matrices by cohort (PPMI/ PDBP, see section 3.2.1 for a description of the cohorts), timepoint and diagnosis ('Case' or 'Control', as defined by the case_control_other_latest variable). Principle components (PCs) were calculated to capture hidden confounders and the largest 8 were used for covariate correction, as these significantly correlated with known covariates derived from the metadata such as 'RIN', 'Box' and 'Plate'. Log10 TPMs were covariate corrected using a linear model with the formula: $tpm \sim age + sex + medication\_status + PC1 + PC2 + ... + PC8$.

It was important to include age, sex and medication status in addition to PC axes because although the PC axes captured RIN well, high PCs did not capture age and sex as strongly. In this work it was important to ensure that these were accounted for because age and sex are known to affect PD pathogenesis. In addition, it was important to ensure cell type proportions were captured by the PCs included in the model as differential proportions between PD cases and healthy controls arising as a result of the disease state could otherwise confound results. Unfortunately, haematological data was not collected for PDBP cohort, however, for the PPMI cohort it was collected for 88% of PPMI participants at the baseline visit (M0). Thus it was not possible to confirm that PCs captured cell type proportions for the exact data being utilised in the present analyses (PDBP-M0 and PPMI-M36). However, this could be confirmed for PPMI-M0, which contains the majority of the same participants sampled at PPMI-M36. Figure 3.4 presents the heatmap for PPMI-M0 healthy control participants showing Spearman correlation coefficients between the proportions of five cell types and the first 10 principle components. PCs four and five in particular capture neutrophils and lymphocytes, the most prevalent cell types providing assurance that cell type proportions are being captured by large PCs. Following fitting of the linear model, residuals were extracted and utilised in down-

stream analyses.

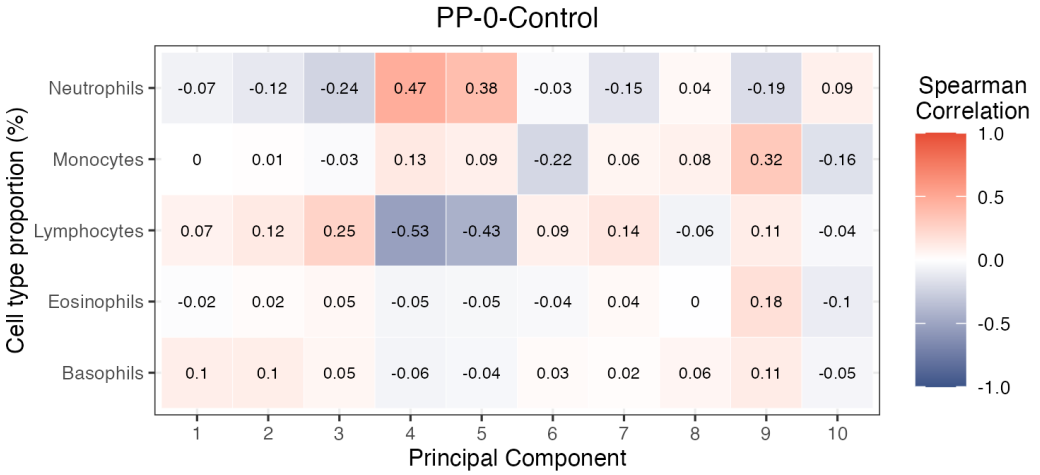

Figure 3.4: Principal component and cell-type proportion correlations for PPMI-M0 healthy control participants

3.2.2.2 Mitochondrial post-transcriptional modification data

Mitochondrial PT modifications were inferred from the RNA-Seq data using bespoke software developed by Alan Hodgkinson[21,19,22]. This inference is based on strand termination events resulting from the physical interference of the reverse transcriptase (RT) enzyme during the cDNA synthesis step in library preparation by the modification which generate random nucleotide incorporation, manifesting as mismatches to the reference genome in the resultant sequencing data. Although the ratio of these mismatch events is not a perfect match to true PT modification rates, the two are highly correlated[19].

To determine mismatch rates from the RNA-Seq data, first, samtools v1.4.1 was used to retain only properly paired and uniquely mapped reads. This was a quality control strategy to ensure that truly mtDNA originating reads were being considered in this step, as opposed to reads arising from NUMTs. Next, allele counts at each mitochondrial position were derived using samtools mpileup, considering only sites with a nucleotide quality score $\geq$ Phred 30 and $\geq$20x coverage. The software was configured to utilise a genome reference file corresponding to build B37, and this was not changed to B38 as the mitochondrial genome remains

unchanged between versions[149]. The pileup files were then used as input into the mismatch proportion calculation step. The proportion of nucleotide mismatches compared to the reference was calculated for each mitochondrial genomic position, calling the presence of a modification if all four nucleotide bases were present in equal proportions for that site, with the rationale that this was indicative of a modification-induced reverse-transcription error being made at this position during the creation and amplification of cDNA. To enable running of this pipeline across all PPMI and PDBP samples using Google Cloud computation via the Broad Institute's Terra platform, a WDL (a language developed for the Broad Institute's genomic analysis pipelines) workflow for these scripts was developed by me (`https://api.firecloud.org/ga4gh/v1/tools/aine_fb_ucl:calculate_mismatch_proportion_workflow/versions/14/plain-WDL/descriptor`). This workflow is publicly available and is configured to run on Terra using a docker image which is also publicly available on DockerHub (`ainefairbrotherbrowne/terra_dockers:5`).

The per-sample mismatch proportion files were subsequently downloaded to the local server and aggregated into a sample-by-position matrix in R. Mismatch proportion data did not undergo covariate correction, unlike the expression data. The reasons for this were four-fold. Firstly, the dependent variable in this case is a ratio, which makes it difficult to accurately model using a linear model -based covariate correction strategy. Secondly, there is missing data, reducing the likelihood of accurately modelling mismatch proportion with respect to multiple covariates. Thirdly, obtaining axes of variation using PCA would be impossible without dealing with the missing data through omission or imputation. A common omission strategy is to perform list-wise deletion, but in this case would result in a large reduction of power resulting from a high sample or mitochondrial position exclusion rate. This would render analyses of smaller sub-groups, in particular, under-powered. Another canonical approach is imputation, but robustly imputing this modification data is potentially very challenging particularly as it is an inferred phenotype without an available gold standard with which to assess the robustness of the output.

Finally, analyses carried out using raw versus covariate corrected expression data resulted in only subtle case-control comparison P-value changes, with distributions and inter-group differences being retained. As such, the modification rates were analysed as raw log10 transformed values ($log10(mismatchratio + 1)$).

3.2.2.3 Metadata

Metadata pertaining to samples and participants is stored in AMP-PD Google Cloud storage in category-, sample ID- (the ID assigned to the biological sample) and participant ID- (the ID assigned to the individual) specific tables. As such, downloading a cohesive metadata table necessitated performing a number of SQL queries to obtain and aggregate the target variables. SQL queries were performed by Ana-Luisa Gil Martinez in a Jupyter Notebook, and run on the Terra platform. The package 'bigrquery' was utilised for integration of Python and BigQuery (BigQuery is a Google data storage 'warehouse' that supports SQL queries). Seven SQL queries were performed to obtain information pertaining to case-control status, demographic, enrolment status, clinically reported genetic status, genetic variants, PD medical history, sample (e.g. visit month) and participant (e.g. age and sex). The output tables yielded were joined by sample ID or participant ID and downloaded to the local server. This metadata was subsequently utilised for dissecting the AMP-PD structure, participant selection, covariate correction, case-control designation and other downstream analyses.

3.2.3 Mean comparisons

Mean comparisons for log10 residual TPMs and log10 mismatch proportions were performed in R. For the expression data which was approximately Gaussian, a two-way student's T-test was applied. For the modification data which was negatively skewed, the non-parametric equivalent, a Wilcoxon signed-rank test, was used. All P-values were Bonferroni corrected, and added to the violin-boxplot visualisations to aid interpretation of inter-group distribution differences.

3.2.4 PCA clustering

Principal component analysis was carried out on expression data derived from 15 mtDNA-encoded OXPHOS genes and 148 nDNA-encoded OXPHOS genes. Participants were then projected onto PC space to observe clustering and uncover hidden patterns in the data not observable through mean comparisons or simple distribution visualisation. The nDNA-encoded OXPHOS gene list was obtained from the Broad Institute MitoCarta3.0 MitoPathways database, called "OXPHOS". PCA was performed in R using the 'stats::prcomp' function. All P-values were Bonferroni corrected.

3.3 Results

3.3.1 Exploring mitochondrial transcriptional profiles in the AMP-PD cohorts

The overarching aim of this chapter is to understand whether the disease state modulates or is driven by changes in the mitochondrial transcriptome. Specifically, this chapter explores whether observable changes exist: (i) Between PD cases (sporadic and monogenic) and healthy controls and, (ii) Between monogenic symptomatic carriers and asymptomatic carriers. The central question of this work considers whether mtDNA gene expression changes across forms of PD, and whether these changes are likely to be part of the underlying disease aetiology or are in fact a consequence of the disease state itself. To look at this, two mitochondrial transcriptional features were analysed: mtDNA encoded gene expression and PT modification of mitochondrial RNA. These two types of feature are interconnected, with PT modifications influencing mitochondrial expression levels in both an upstream and downstream manner. The following sections describe these transcriptional features in detail (sections 3.3.1.1 and 3.3.1.2).

3.3.1.1 Transcriptional feature I: post-transcriptional modification

Mitochondrial transcription (as detailed in section 1.1.2) results in two main (heavy and light chain -derived) near-genome length polycistrons. Post-transcriptional (PT) modification of the 9th position of tRNAs (p9 sites) on the polycistron confers transcriptional flexibility, allowing divergence from the 1:1 processed mRNA ratio that might be expected to arise from polycistronic transcription. Certain N1-methyladenosine (m1A) and N1-methylguanine (m1G) modifications are readily detected in RNA-Seq data (see section 1.1.2 for further details) and as such, 21 m1A/G modification sites of known functional importance derived from Hodgkinson et al., (2014) were considered in the first instance (see table 3.1 for details of the 21 sites)[22]. Of these 21 sites, 19 were tRNA p9 sites, and the remaining two were in *MT-RNR2* and *MT-ND5*. Considering modifications at tRNA p9 sites, ribosomal gene sites and respiratory chain gene sites allowed surveillance of modifications associated with a variety of downstream consequences, including tRNA stability and altered mitoribosome interaction.

Using the AMP-PD data, inference of modification rates at 21 sites was performed for each cohort (PPMI and PDBP) at all timepoints, yielding modification rates for between 470-1448 participants per-timepoint in PDBP and 540-873 participants per-timepoint in PPMI. It was important to consider the sparsity of the modification data which is inherently limited by the methodology used to detect the modifications from RNA-seq data. For further details of the modification inference pipeline see section 3.2.2.2, but in brief, a minimum coverage of 20 reads is required to call the presence or absence of a modification, and as such, 'NA' values are used to mask out data points with insufficient coverage. Additionally, although modifications have previously been observed at each of the 21 sites analysed, in many cases, sites have a zero value assigned to them, indicating that no modification was detected in that individual. There are a few possible explanations for this. There may be genuinely no modification for this individual at this site, in which case a zero value confers valuable information. Within this, it may be that no modifica-

mtDNA position (bp)	Gene	Description
585	MT-TF	9th position of tRNA
1610	MT-TV	9th position of tRNA
3238	MT-TL1	9th position of tRNA
4271	MT-TI	9th position of tRNA
4392	MT-TQ	9th position of tRNA
5520	MT-TW	9th position of tRNA
5647	MT-TA	9th position of tRNA
5721	MT-TN	9th position of tRNA
5818	MT-TC	9th position of tRNA
5883	MT-TY	9th position of tRNA
7526	MT-TD	9th position of tRNA
8303	MT-TK	9th position of tRNA
9999	MT-TG	9th position of tRNA
10413	MT-TR	9th position of tRNA
12146	MT-TH	9th position of tRNA
12274	MT-TL2	9th position of tRNA
14734	MT-TE	9th position of tRNA
15896	MT-TT	9th position of tRNA
15948	MT-TP	9th position of tRNA
2617	MT-RNR2	MT-RNR2 gene
13710	MT-ND5	MT-ND5 gene

Table 3.1: Table describing the 21 mitochondrial modifications initially considered in this study.

tion was present in the temporal state captured at the point of RNA extraction, or perhaps the modification is tissue-specific, and as such it is not possible to detect it in blood. However, zero values may also be non-informative, indicating low coverage over the site. High levels of zeroes and missingness would limit statistical power downstream, and as such it was necessary to evaluate both the missingness and information content at each site and to decide which sites could be included in downstream analyses.

Figure 3.5a shows, for each position, 'missingness', defined as % of samples with a non- 'NA' modification rate plotted against 'information'. 'Information' is defined on the y-axis as % of samples with modification rate above zero, and is represented as the median modification rate and visualised by point size. Site 2617 has high information and no missingness, wherein 0% of samples have missing data and 100% of samples have non-zero values, coupled with a high cross-sample median modification rate (PDBP, mean=0.405, min=0.32, max=0.837; PPMI, mean=0.444, min=0.114, max=0.857). This is consistent with previous work which found

that the inferred methylation at 2617 in CARTaGENE whole blood samples was between 0.48 and 0.72[21]. In contrast, site 3238 has low information (PDBP, mean=0.000390, min=0, max=0.037; PPMI, mean=0.000614, min=0, max=0.05) and missingness of approximately 80% across cohorts. Again, this is in line with previous work showing low modification rates (mean of 0.009 across multiple body tissues) at this position[21].

To balance surveying a range of mitochondrial modification sites and retaining sites with sufficient statistical power, a cut-off of 25% was applied for missingness and information. To be carried forward, a site had to have >25% of samples with a modification rate >0 and not 'NA', and had to satisfy these criteria in both PPMI and PDBP cohorts. The seven sites satisfying this (2617, 5647, 5721, 5818, 5883, 7526, 9999) are coloured blue in figure 3.5a. Their positions in the context of the mitochondrial genome are shown in figure 3.5b, where 2617 is in an mtDNA-encoded ribosomal gene (*MT-RNR2*) and the other six are at p9 sites. In this study, missingness was surveyed across cohorts only for simplicity. However, it is important to acknowledge that missingness levels may also differ between cases and controls and that this may be biologically meaningful. As such, the assessment here could be improved by surveying missingness not only across cohorts, but also across cases and controls to ensure that bias is not being introduced into the analysis by selecting methylation sites with lower missingness in controls compared to cases, for example.

For each of the seven positions selected, the log10-transformed distributions of their modification rates are shown in figure 3.5c. Overall, distributions of modification rates were highly negatively skewed and bimodal, wherein a large peak at zero was followed by a smaller peak at a higher rate. This was true across cohorts and timepoints for all sites except the *MT-RNR2* site 2617. 2617 appeared to have a biomodal distribution, particularly in the PPMI cohort, but never had a rate of 0 and was consistently modified at a higher frequency than all other sites across cohorts and timepoints. The other six sites grouped approximately by genomic proximity, where sites 5647, 5721, 5818 and 5883 displayed similar modification frequency

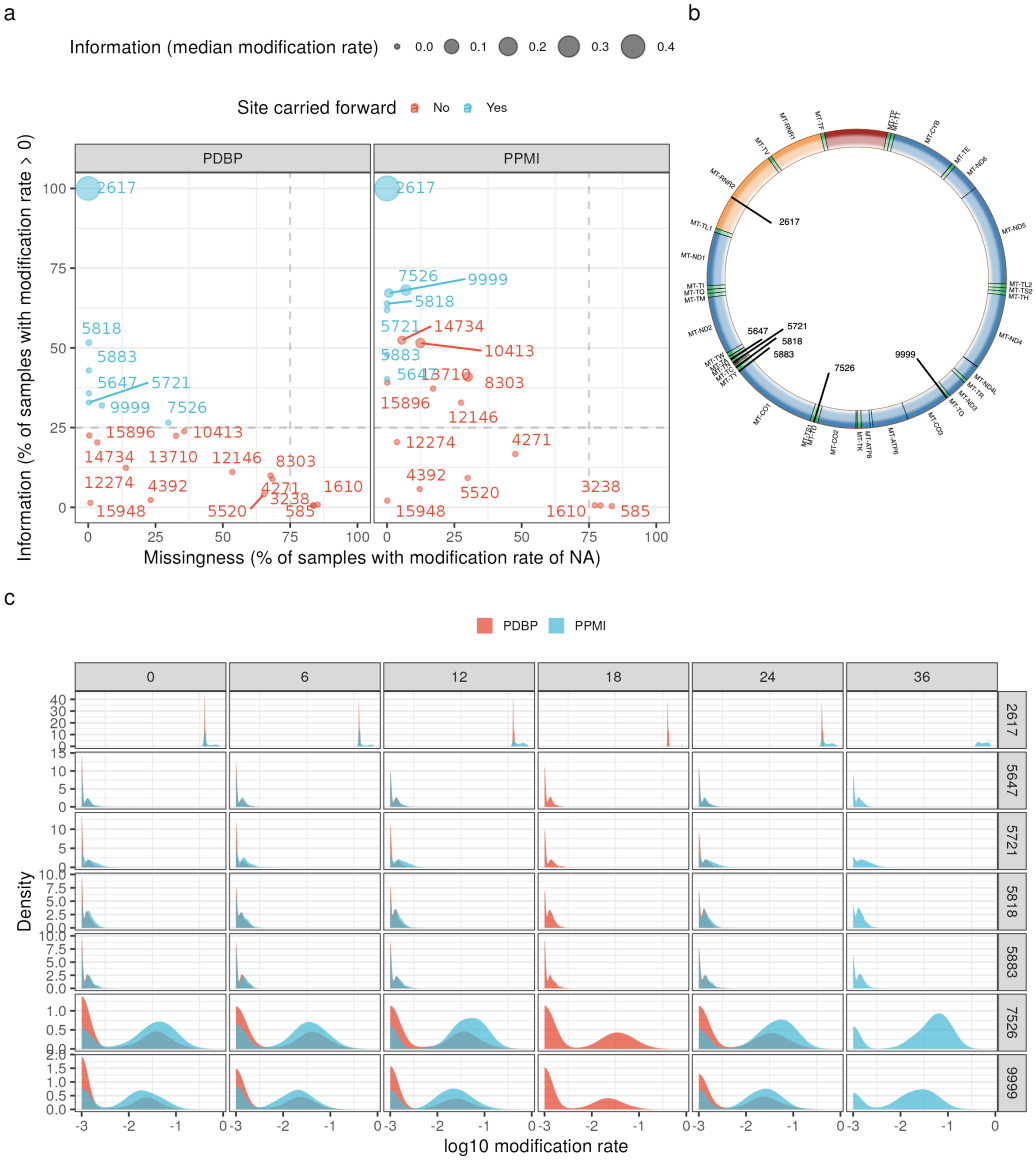

Figure 3.5: Exploration and selection of 21 mitochondrial modification sites. A. Cohort-faceted dot-plot to show, for 21 mitochondrial modification sites, 'missingness' (% of samples with modification rate of 'NA') against 'information' (% of samples with modification rate >0). Point size represents the median modification rate across all partipants and all timepoints. B. Schematic showing the mitochondrial genome, where tRNA genes are indicated in green, ribosomal genes in orange and respiratory chain components in blue. Black lines labelled with their base positions indicate the seven modifications sites forward for further analyses. C. Log10 transformed modification rates, faceted by position and timepoint for the PPMI (blue) and PDBP (red) cohorts.

profiles across cohorts and timepoints, as did sites 7526 and 9999.

3.3.1.2 Transcriptional feature II: gene expression

The second mitochondrial transcription feature analysed was expression arising from all 15 mtDNA-encoded genes: *MT-ND1*, *MT-ND2*, *MT-ND3*, *MT-ND4L*, *MT-ND4*, *MT-ND5*, *MT-ND6*, *MT-CO2*, *MT-CO3*, *MT-ATP8*, *MT-CYB*, *MT-CO1*, *MT-ATP6*, *MT-RNR1* and *MT-RNR2*. In contrast with the modification rate data, these data were complete due to the very high expression levels of mtDNA genes. As such no missing data management protocol was required and all 15 genes were included in downstream analyses.

For each of the 15 genes, the log10-transformed distributions of residual TPM values are shown in figure 3.6. *MT-CO1* displayed the highest median expression of all the mtDNA genes, whilst *MT-ND6* had the lowest median expression, but, uniquely, had a long tail containing higher expression values. Expression distributions across all genes, cohorts and timepoints approximated the normal distribution. In general, genes had similar distributions between PPMI and PDBP, with the exception of the mitochondrial ribosomal genes, *MT-RNR1* and *MT-RNR2*, where for both genes, median expression was markedly higher in the PDBP cohort than in the PPMI cohort.

3.3.1.3 Case-control differences in mitochondrial transcriptional signatures

In studying transcriptional profiles across forms of PD, this work did not consider how these profiles changed or progressed across clinical visits. As such, the data was made static by selecting a single timepoint, thus utilising unique participants, making comparisons within-timepoint and removing any timepoint-specific effects. Both mitochondrial expression (Fig. 3.6) and modification rate (Fig. 3.5c) displayed relative consistency across timepoints, and as such, for timepoint selection, sample number and cross-cohort participant profile matching were considered the primary selection criteria. For PDBP, the most optimal timepoint was M0 because considerably more participants were sampled at M0 than at any other timepoint, of-

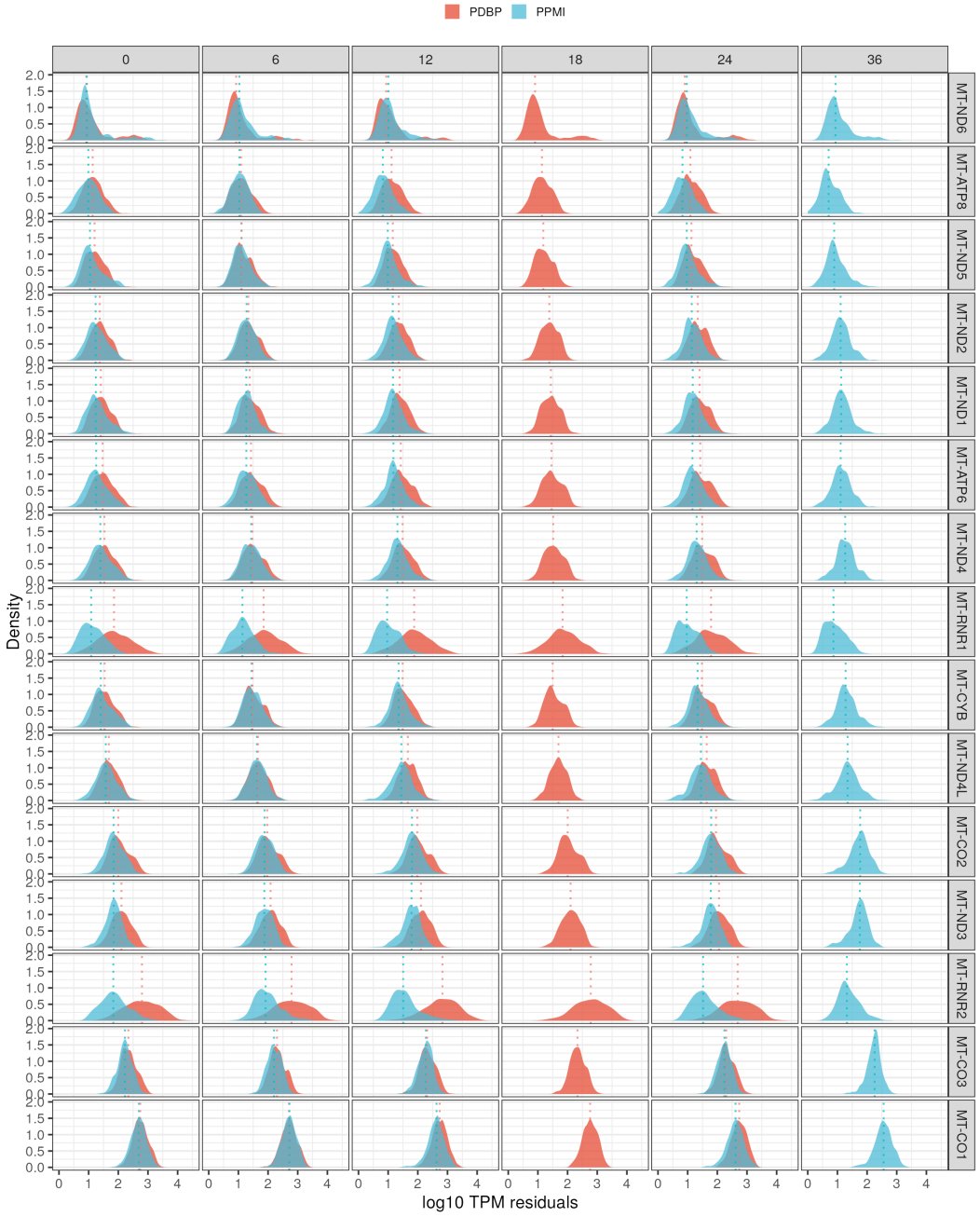

Figure 3.6: Exploration of mtDNA-encoded gene expression levels. Density plots to show log10-transformed TPM values for 15 mtDNA-encoded genes, faceted by position and timepoint for the PPMI (blue) and PDBP (red) cohorts. Vertical dotted lines represent the median.

fering maximum statistical power (Fig. 3.2c). For PPMI, the most optimal timepoint was less evident due to inconsistent participant category (Fig. 3.3c) and medication status (Fig. 3.2d) composition across timepoints. Most important for downstream analyses was the availability of 'Genetic' and healthy control participants, for which M0 and M6 were unsuitable due to lacking healthy control participants. M36 had very few 'Genetic' participants, leaving M24 and M12 which were both reasonable matches for the participant composition (category and medication -wise) of PBDP M0. M12 was selected due to its marginally larger numbers of healthy control and 'Genetic' participants. Matching timepoints between cohorts was initially considered, but cohort-specific recruitment criteria meant that timepoints, with respect to age and disease stage, were non-comparable across cohorts. For these reasons, work in this chapter and the next will largely consider PDBP-M0 and PPMI-M12 participants.

The first aim was to understand whether transcriptional phenotypes differed between PD cases and controls. To this end, for each phenotype (expression from 15 mtDNA-encoded genes and modification rate at seven mitochondrial sites), distribution differences between healthy control, genetic cohort (available only in PPMI) and sporadic PD participant categories were assessed. Expression values were corrected for the effects of age, sex, medication status and hidden confounders using linear regression (see section 3.2.2.1 for methodological details). Modification rates were not corrected for covariates but were instead analysed as log10 transformed raw values (see section 3.2.2.2 for methodological details).

Using a non-parametric test to assess case-control differences in the negatively skewed modification rate distributions, no significant differences between healthy control and sporadic PD participants were observed at any modification sites (P-adjusted$>$0.6 across all sites). This was the case for both PPMI and PDBP cohorts (Fig. 3.7). However, highly significant modification rate differences were observed between healthy control and genetic cohort PD participants in the PPMI cohort at sites 2617 (P-adjusted=6.06e-21), 5721 (P-adjusted=1.12e-03) and 7526 (P-adjusted=3.36e-05) (Fig. 3.7). At all three sites, the genetic cohort PD partici-

pants displayed lower modification rates than the healthy control participants.

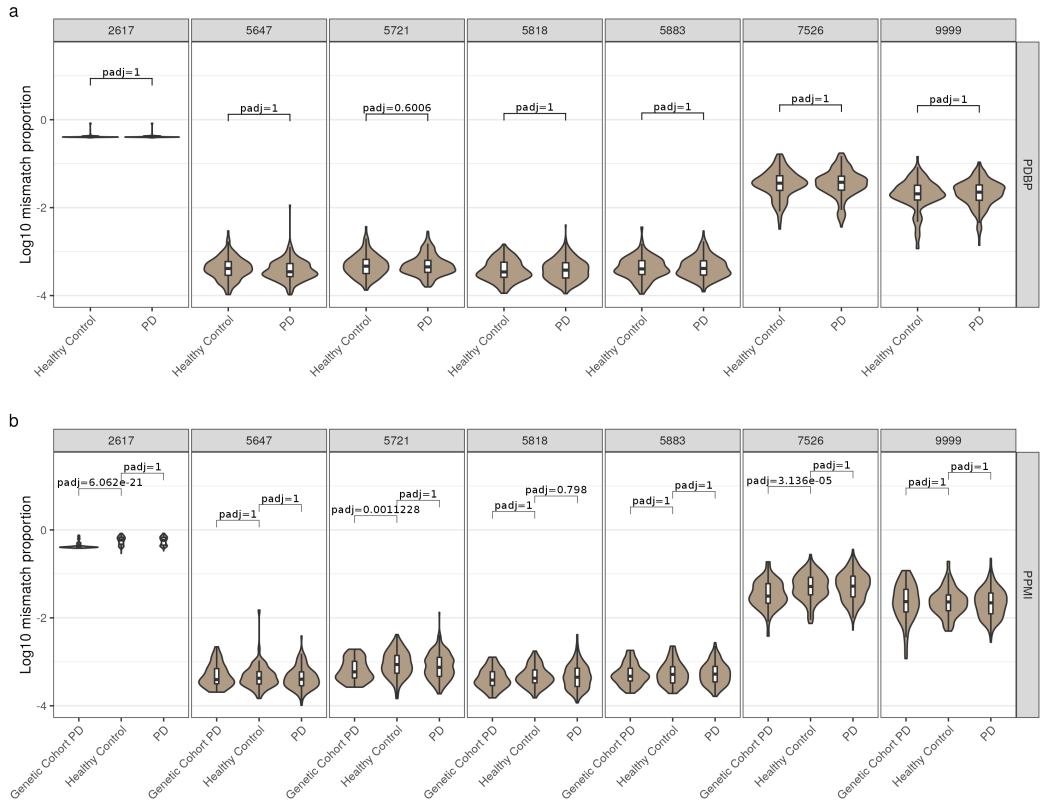

Figure 3.7: Distributions of log10 transformed modification rates at 7 mitochondrial sites in: A. The PDBP cohort and B. The PPMI cohort.

Using a parametric test to assess case-control differences in expression rate, no significant differences were found between healthy control participants and sporadic PD participants in the PDBP cohort (Fig. 3.9). In the PPMI cohort, however, significant differences between healthy control participants and sporadic PD participants were detected for 10 genes. For all 10 genes, expression in sporadic PD participants was higher than in healthy control participants. Differences were more striking, however, between genetic cohort and healthy control participants (in PPMI), with highly significant differences detected in the expression of 12 genes. For 11/12 genes, expression in genetic cohort participants was higher than in healthy control participants, however for 1/12 (MT-CO3), the pattern was reversed.

The central trend was that genetic cohort PD participants had significantly higher mtDNA-encoded gene expression than healthy control participants across

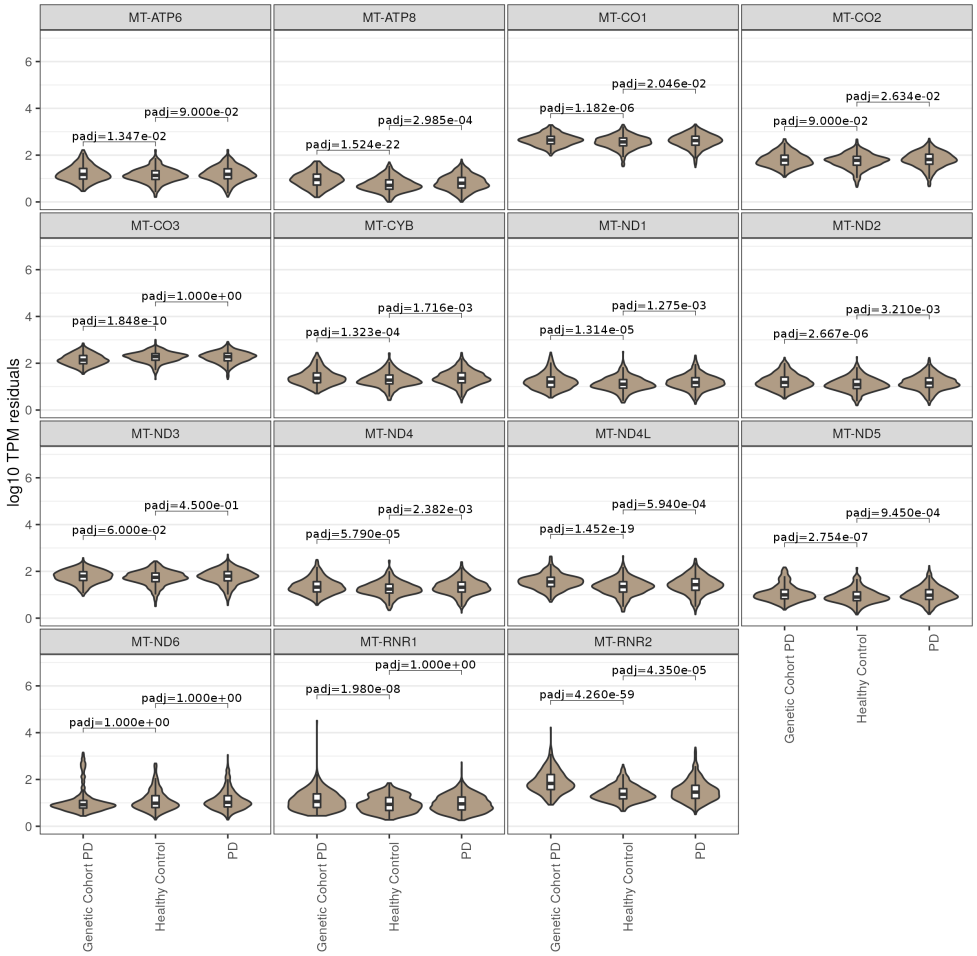

Figure 3.8: Distributions of gene expression (log10 TPMs) for 15 mtDNA-encoded genes in the PPMI cohort (visit month 12). P-value annotations are the output of t-tests between healthy control participants and all other groups. All P-values have been Bonferroni corrected.

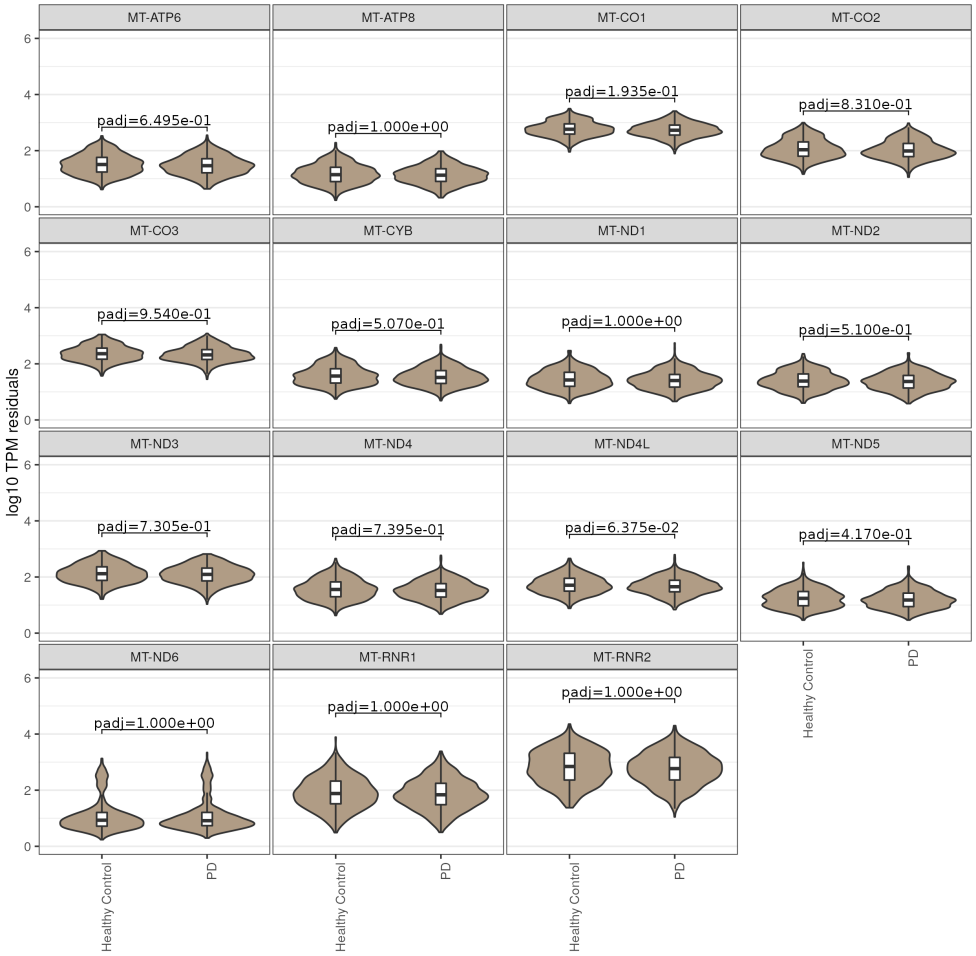

Figure 3.9: Distributions of gene expression (log10 TPMs) for 15 mtDNA-encoded genes in the PDBP cohort (visit month 0). P-value annotations are the output of t-tests between healthy control and sporadic PD participants. All P-values have been Bonferroni corrected.

the mitochondrial genome. Coupled with this, 3/7 mitochondrial genomic sites displayed lower methylation rates in genetic cohort PD compared with healthy control participants. The sporadic PD participants also showed significantly higher mitochondrial expression than healthy control participants, but only in the PPMI cohort. This could not be replicated in the PDBP cohort. Additionally, no changes in mitochondrial modification could be observed in sporadic PD participants, a pattern that was consistent across both cohorts.

Although these findings were interesting and provided a broad overview of case-control differences in mitochondrial transcriptional phenotypes, this analysis assessed inter-category differences using simple mean comparisons and did not consider the presence of sub-groups of participants. To more robustly assess differences in case-control transcriptional signatures, four key improvements were made. The first improvement was to utilise a dimensionality reduction method. Principal component analysis (PCA) is a non-parametric method wherein, the dimensionality of the data is reduced, allowing inter-group differences and similarities to be assessed across multiple variables simultaneously. This method is well-suited for a complex dataset such as this, which has multiple dimensions (e.g. multiple phenotypes, groups, participants) and may have hidden structure that is not readily apparent using simpler statistical and visualisation methods. Employing PCA, the aim was to uncover hidden structure and further understand grouping and relationships within the data. The second improvement was to advance the analysis by interrogating transcriptional profile differences between sub-groups of the genetic cohort PD participants as defined by gene variants. This was of interest because is not currently known how mitochondrial transcriptional profiles differ between PD cases associated with differential molecular aetiologies. This has potential implications for understanding the underlying mechanisms of PD. To this end, the genetic cohort PD group was split into three groups according to the monogenic variants carried: *SNCA*, *GBA* and *LRRK2*. The third improvement was to increase sample numbers in PPMI participant sub-groups by adding M24 orphan participants. In brief, these were participants with data collected at M24 but not M12, and as such were added

to the timepoint M12 participants (see section 3.2.1.4 for further details and figure 3.3b to see where participant number gains were made). The fourth improvement was to include the complete set of 163 OXPHOS genes — mtDNA and nDNA - encoded — as opposed to mtDNA-encoded genes only. This was firstly to assay a coherent mitochondrial process and secondly to provide power for dimensionality reduction.

Finally, it is important to note that the modification rate data was excluded from the following analyses due to the presence of missing data (see methods section 3.2.2.2 for further details). As such, the dimensionality reduction analyses as described in section 3.3.1.4 were performed using the expression data only.

3.3.1.4 Clustering AMP-PD participants on mitochondrial expression levels

Clustering was performed to observe hidden patterns and clusters that may exist among participants in terms of OXPHOS expression profiles. PCA was performed on the expression values of 15 mtDNA-encoded genes in addition to expression values of 148 nDNA-encoded OXPHOS sub-unit genes. A gene list for the latter was downloaded from the Broad Institute's MitoPathways3.0 resource[75]. The nDNA genes encode the complementary proteins that, together with the mtDNA-encoded genes, encode all core components of the OXPHOS pathway.

Figure 3.10a through 3.10i show the projection of participants onto PC space, where the first 10 PC axes together explain 82.7% of the variance in the expression data (3.10j). Nine two-dimensional spaces generated by sequential combinations of these PC axes have been visualised here.

The most obvious separation of participants occurs at PC7 (explaining 3.95% of total variance), which generates a cluster of genetic cohort samples of all three gene sub-groups that is clearly separate from the sporadic PD and healthy control participants. This observation was supported empirically by correlations between PC eigenvectors and PD gene status, yielding a highly significant Spearman's ρ of 0.51 (P-adjusted=3.975e-45) between *LRRK2* status and PC7 (Fig. 3.10k). *SNCA* (ρ=0.18, P-adjusted=3.590e-06) and *GBA* (ρ=0.14, P-adjusted=1.155e-03)

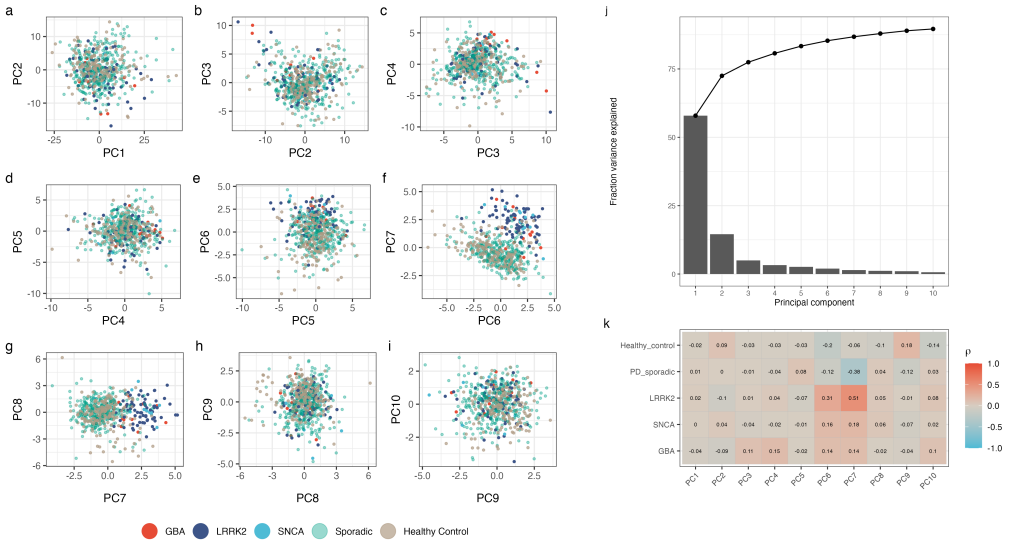

Figure 3.10: PCA clustering of PPMI participants on OXPHOS expression. A-I. Scatter plots to show PPMI participants (*GBA*: n=16, *LRRK2*: n=80, *SNCA*: n=12, sporadic PD: n=336, healthy control: n=170) projected onto PC space, where PCs were calculated using the expression of OXPHOS genes. J. Scree plot to show the percent of total variance explained by each PC. K. Heatmap to show the correlation between participant diagnoses and each PC. The colour scale and labels on cells indicate the Spearman correlation coefficient (ρ).

status were also significantly correlated with PC7. Healthy control (ρ=-0.06, P-adjusted=6.350e-07) and PD (ρ=-0.38, P-adjusted=3.435e-10) status were significantly correlated with PC7, but with opposite directionality, confirming the visually observed separation of *GBA*, *LRRK2* and *SNCA* participants from healthy control and sporadic PD in PC space. Evidence of separation between sporadic PD and healthy control participants was observed at PC9 (Fig. 3.10h-i), which is significantly associated, though with opposing directionality, with both healthy control status (ρ=0.18, P-adjusted=5.98e-05) and sporadic PD status (ρ=-0.12, P-adjusted=8.71e-03). While the visual separation between the sporadic PD and healthy control participants is difficult to discern, the statistical assessment points to separation of these groups.

The sporadic PD and healthy control participant separation observed in the PPMI cohort was replicated in the PDBP cohort (Fig. 3.11). Healthy control (ρ=-0.22, P-adjusted=1.656e-13) and sporadic PD (ρ=-0.22, P-adjusted=1.656e-13) status were significantly correlated with PC2 and with opposite directionality of cor-

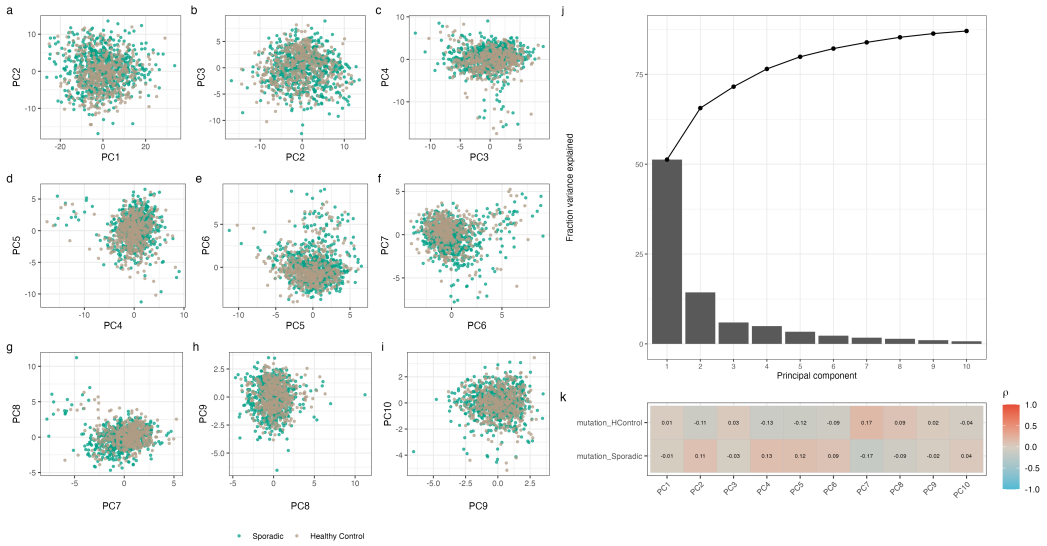

Figure 3.11: PCA clustering of PDBP participants on OXPHOS expression. A-I. Scatter plots to show PDBP participants (sporadic PD: n=688, healthy control: n=492) projected onto PC space, where PCs were calculated using the expression of OXPHOS genes. J. Scree plot to show the percent of total variance explained by each PC. K. Heatmap to show the correlation between participant diagnoses and each PC. The colour scale and labels on cells indicate the Spearman correlation coefficient (ρ).

relation coefficients, thus confirming the visually observed separation of healthy control and sporadic PD participants. Replication of the genetic cohort participant clustering was not possible in the PDBP dataset, as it only contains sporadic PD and healthy control participants.

In summary, changes in OXPHOS gene expression were observed in PD participants compared to healthy controls, regardless of the underlying genetic cause (i.e. monogenic or sporadic), though this was particularly striking in the monogenic forms. As such, it was found that the expression of genes comprising a key mitochondrial process, OXPHOS, was altered in the disease state.

In the PPMI dataset, in addition to the genetic cohort PD participant category — comprised of symptomatic individuals with confirmed pathogenic variants in *LRRK2*, *GBA* or *SNCA* — there exists a 'genetic cohort unaffected' category. These are asymptomatic PD variant carriers, meaning that they have not yet developed, or may not develop, PD symptoms. The presence of monogenic symptomatic and asymptomatic participants presented a unique opportunity within which to compare

the transcriptional profiles between two defined disease stages. Here, the aim was to utilise these groups to provide insights into the timing of the shifts that have been observed between cases and controls. To this end, clustering was performed on OXPHOS gene expression in the same way as previously described, but using three participant groups: 'Genetic diagnosis - affected' (n=111), 'Genetic diagnosis - unaffected' (n=87) and healthy control (n=170) (Fig. 3.12). PC6 clearly separated the healthy control participants from the 'Genetic diagnosis - affected' and 'Genetic diagnosis - unaffected' participants with the latter two groups remaining clustered together (Fig. 3.12e and f). Healthy control status (ρ=0.66, P-adjusted=3.180e-46), 'Genetic diagnosis - affected' status, (ρ=-0.54, P-adjusted=9.990e-28) and 'Genetic diagnosis - unaffected' status (ρ=-0.19, P-adjusted=5.670e-03) were significantly correlated with PC6 (Fig. 3.12k). Crucially, 'Genetic diagnosis - affected' and 'Genetic diagnosis - unaffected' status had opposing directionality of correlation coefficients with PC6 (positive) compared to healthy control status (negative), confirming the visually observed separation of healthy control and sporadic PD in PC space. The clustering of affected and unaffected monogenic PD cases and separation from healthy controls suggests that the observed shifts in OXPHOS expression are present in asymptomatic cases as well as symptomatic cases. This provides a novel insight, pointing to the idea that changes in mitochondrial processes may actually occur before symptom onset.

3.3.1.5 Validating PCA clustering

To test whether the OXPHOS gene set was exerting a gene set -specific effect participant separation and that separation was not occuring by chance, comparison to a null distribution was performed. This distribution was generated by running 5000 iterations of PCA, each time selecting a random gene set from the set of all genes expressed, matching the set size to that of the OXPHOS gene set (163). The first 10 PC axes were extracted from each iteration, correlating each with participant diagnosis and yielding a Spearman's ρ and corresponding P-value. For each PC-diagnosis pair, the null distributions (5000 PC-diagnosis Spearman's ρ and P-values) were visualised, superimposing the real value generated using the OXPHOS gene set. An

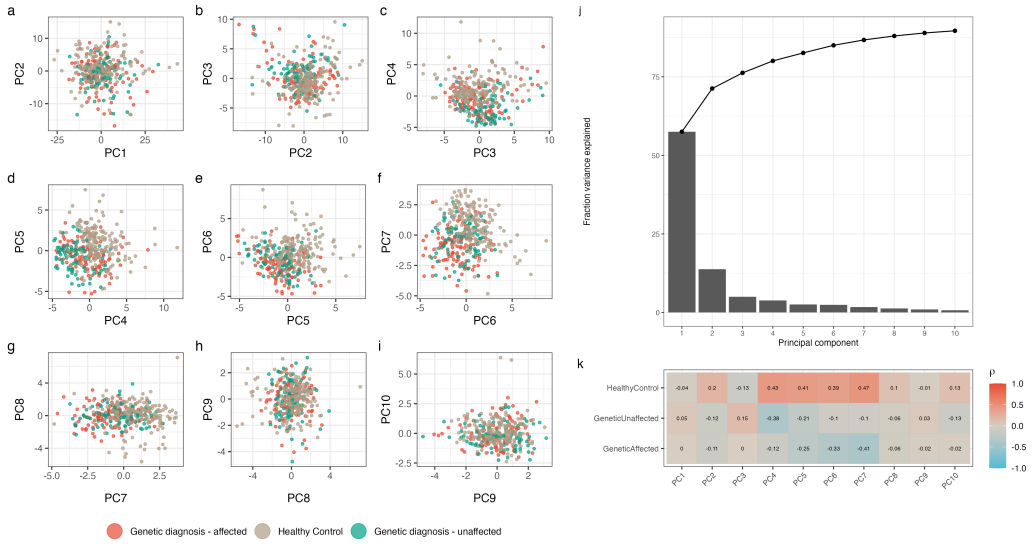

Figure 3.12: PCA clustering of PPMI participants (symptomatic/ asymptomatic/ healthy control) on OXPHOS expression. A-I. Scatter plots to show PPMI participants ('Genetic diagnosis - affected': n=111; 'Genetic diagnosis - unaffected': n=87; healthy control: n=170) projected onto PC space, where PCs were calculated using the expression of OXPHOS genes. J. Scree plot to show the percent of total variance explained by each PC. K. Heatmap to show the correlation between participant diagnoses and each PC. The colour scale and labels on cells indicate the Spearman correlation coefficient (ρ).

real value more extreme than the null distribution indicated that that the OXPHOS gene set separated the participants more often than would be expected by chance.

Figure 3.13 shows the output of this bootstraping analysis for the clustering of *SNCA*, *LRRK2*, *GBA*, sporadic PD and healthy control participants (shown in Fig. 3.10), displaying Spearman's ρ P-values in 3.13a) and Spearman's ρ values in(3.13b). It is clear that the correlations with PC7 for all diagnoses (bar healthy control) are markedly more extreme than the null distribution. There is also evidence for the opposing correlation directionality of healthy control and sporadic PD diagnosis with PC9 being more extreme than the null distribution, though this is more subtle. This points to the 163 mtDNA and nDNA -encoded OXPHOS gene set as exerting a specific and likely biological meaningful effect on the clustering of participants.

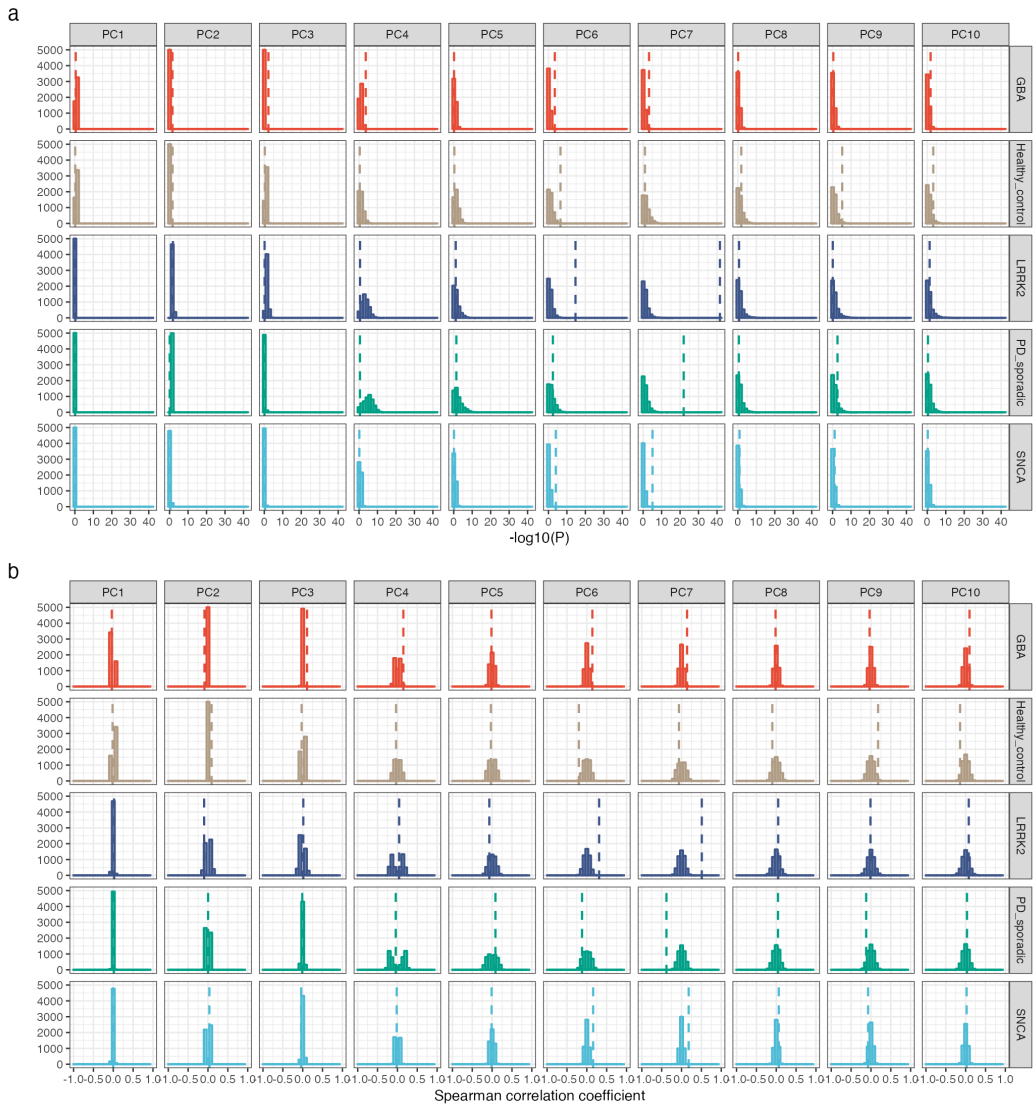

Figure 3.13: Validating participant clustering of *SNCA*, *LRRK2*, *GBA*, sporadic PD and healthy control participants using PCA bootstrapping. A. Null distributions for each PC-diagnosis pair containing 5000 Spearman's ρ P-values that were generated using random gene sets. B. Null distributions for each PC-diagnosis pair containing 5000 Spearman's ρ that were generated using random gene sets.

3.4 Discussion

The aetiology of PD is not fully understood, but mitochondrial abnormalities have long been postulated to contribute to disease pathogenesis. Indeed, in their essential role in energy production makes normal mitochondrial function key to the integrity of dopaminergic neurons. Pathogenic variants in several monogenic PD-related genes have been found to be involved in mitochondrial dysfunction through a variety of mechanisms. This work focuses on *LRRK2*, *SNCA* and *GBA* due to availability of sizeable variant carrier cohorts via the AMP-PD project.

Although work to date on monogenic PD offers a nuanced and detailed mechanistic understanding of certain aspects of pathogenesis, it largely utilises non-human disease models such as mice[150] and zebrafish[151] or human *in vitro* cell models such as induced pluripotent cell (iPSC) lines[152] and fibroblast cell lines[153,154,155]. While these studies are essential for unpicking disease mechanisms, they are rarely suitably powered for population-level analyses seeking to reveal overall patterns, prevalence and distribution of disease features. Indeed, understanding these overall patterns may provide insight into currently lesser understood mechanisms of sporadic PD. This is highly relevant in the context of present research focuses in PD which seek to understand the phenotypic heterogeneity within PD thought to be a manifestation of disease subtypes[144]. Understanding the contribution and relevance of different molecular mechanisms to subgroups of PD patients has implications for biomarker identification, drug development and patient management.

This research aimed to bridge this gap in current knowledge by performing a large-scale, dual-phenotype, direct comparison of mitochondrial transcriptional profiles between monogenic PD cases (with *LRRK2*, *GBA*, and *SNCA* variants), sporadic PD cases and healthy controls. Two central questions were posed. The first asked whether differences in the mitochondrial transcriptome can be observed between monogenic PD cases, sporadic PD cases and healthy control individuals. The second asked whether any differences detected were present prior to, or after symptom onset. Overall, these questions aimed to further understand the role of mitochondrial dysfunction in the pathophysiology of PD.

3.4.1 PD case-control mtDNA-encoded gene expression changes

Marked differences in mitochondrial transcription were detected between monogenic PD cases and healthy controls, including significant expression differences in 12/15 mtDNA genes (Fig. 3.8). Surprisingly, consistently higher expression was observed in monogenic PD participants (symptomatic) compared with healthy control participants across all 12 genes. This was supported by the clustering results, which revealed that monogenic PD participants (symptomatic) clustered separately in principle component space from healthy control participants (Fig. 3.8). Very little quantification of mtDNA-encoded gene expression has previously been performed in a PD case-control paradigm, and none in whole blood from human PD patients. As such there is little opportunity for direct replication of these results. However, a number of possible mechanisms for these observations are discussed henceforth.

Firstly, there is the potential that differential cell type composition between the disease and control state (i.e. varying leukocyte count) could be stochastically confounding these results[156]. However, this is unlikely as a covariate correction strategy was applied that included PCs to capture hidden confounders, such as cell type. Although cell type proportion information was only available for a proportion of 'M0' samples and not the samples utilised in the present analyses, PCs were assessed against these nonetheless (see methods section 3.2.2.1 for details and visualisation). It was found that the PCs included in the model did capture cell type proportions (in the M0 data) and as such, making it unlikely that cell type proportions would be driving the patterns observed here.

Secondly, mitochondrial biogenesis — the process by which new mitochondria are synthesised — triggered by an accumulation of dysfunctional mitochondria may be contributing to higher mitochondrial expression in the disease state. A possible route to this is through increased dysfunction of the mitochondrial population leading to impairment of overall ATP synthesis and reduced ATP availability. In this scenario, an increase in AMP/ATP ratios may enhance the activity of AMP-activated protein kinase (AMPK), a biosensor known to promote mitochondrial biogenesis through the transcriptional regulation of relevant genes[157,158]. Although there is

limited data on mtDNA gene expression in the blood of PD patients, one study observed that peripheral blood mononucleate cells (PBMCs) of individuals diagnosed with PD displayed increased respiratory capacity and elevated ATP production[159]. Interestingly, in a rodent *in vitro* cell model of α-synucleinopathy, it was shown that over-expression of PGC-1α, the so-called 'master regulator' of mitochondrial function, resulted in the up-regulation of nDNA-encoded OXPHOS subunits and that this alleviated α-synucleinopathy neurotoxicity and suppressed dopaminergic neuron loss[160]. Although the Zheng et al., (2010) study deals with nDNA-encoded OXPHOS components only, these results, and those of the present study may point to a mitochondrial biogenesis response[160]. Whether this is arising from pathogenic changes in nuclear-encoded regulatory mechanisms, or perhaps acting in a compensatory manner in response to bioenergetic deficiencies is not clear.

A third potential route to the mtDNA gene expression changes observed in PD cases may be through increased mtDNA content resulting from increased mitochondrial copy number (mtDNAcn). Although, previous work has shown that *reduced* mtDNAcn in blood is a hallmark of PD, a finding that has been replicated in other NDs for which mitochondrial dysfunction has been implicated[161]. As such, it is unlikely that mtDNAcn changes are producing the increase in expression observed here. Additionally, as previously discussed, the expression data was corrected for a number of covariates, including PCs which were shown to correlate to blood cell type proportions. mtDNA copy nunber is known to vary according to cell type[162,163,164], and as such it is likely that at least a proportion of variation in mtDNAcn was captured and corrected for using this method. However, this postualtion does highlight a potential interesting direction for future work involving the estimation and analysis of mtDNAcn in the AMP-PD data[165].

A fourth explanation may involve an increase in mtRNA content (without accompanying raise in mtDNA content) through increased transcriptional rate. Indeed, a mechanism resulting in increased mtDNA expression has been identified, albeit in a non-neurological context. A 2019 study[166] screening for proteins involved in the expression of mtDNA under stress conditions identified Mitochon-

drial Transcription Rescue Factor 1 (MTRES1). MTRES1 was found to be elevated in cells under stress conditions, preventing mtDNA-encoded transcript loss by increasing mtDNA transcription[166]. The authors surmised that this was a protective mechanism against mtDNAcn decline, working to oppose mtRNA loss under conditions of cellular stress[166]. This study presents convincing evidence for mechanisms responsible for modulating the mitochondrial transcriptional rate, pointing to the possibility that this mechanism may explain the PD case expression increase observed in the present study.

Finally, the results observed here might be explained by impaired mitophagy. Variants in all three monogenic PD genes (*GBA*, *LRRK2* and *SNCA*) have been associated with dysregulated mitophagy, the process by which dysfunctional mitochondria are destroyed[167,154,168,152,169]. Additionally, loss-of-function variants in the *PINK1* and *PRKN* genes - encoding key regulators of mitophagy - are known to cause early onset PD. These facts make a compelling case for the role of defective mitophagy in PD pathogenesis. As such, it is important to consider whether the results of the present study are capturing this process. The present analysis includes *SNCA*, *GBA* and *LRRK2* variant carriers, for which evidence to date suggests that the precise route towards defective mitophagy may differ. For example, in *SNCA* variant carriers, accumulation of α-synuclein within mitochondria has been found to increase the rate of mitophagy, likely driven by damage to complex I, mitochondrial fragmentation and energy deprivation[170,171,172]. Given that mitochondrial homeostasis relies on coordinating the removal of damaged mitochondria and generating new mitochondria, perhaps increased mitophagy is driving a matched increase in the expression of mtDNA genes, which may be captured in the present study. However, in *LRRK2* variant (G2019S) carriers, basal mitophagy was seen to decrease rather than increase[173]. In the brains of *LRRK2* (G2019S) mice, expression levels of mitochondrial biogenesis regulators PGC-1β and PGC-1α were increased compared to wild-type (WT) mice, which the authors suggest as indicative of mitochondrial biogenesis activation. Similarly, in PD patients carrying the L444P *GBA* variant, inhibition of mitophagy has been observed alongside increased

mitochondrial content and mitochondrial stress[168]. Taken together, evidence for the differential effects on mitophagy of PD variants points to the idea that dysfunctional mitophagy, regardless of directionality may trigger cellular changes aiming to restore homeostasis. As such, the increased expression observed in monogenic PD cases here may capturing this.

3.4.2 Clustering of participants on (nDNA and mtDNA - encoded) OXPHOS gene expression

The clustering experiments revealed clustering of *SNCA*, *GBA* and *LRRK2* variant carriers, meaning that regardless of the specific gene variant, OXPHOS expression profiles were similar between participants. While this could point to a common increase in OXPHOS expression across all forms of monogenic PD, it might be that a higher resolution approach would reveal differences in OXPHOS expression profiles between groups. This would require increased statistical power through an expansion of participant numbers which is challenging. In fact there is already some evidence to support the view that genetic cohort PD participants with *SNCA*, *LRRK2* and *GBA* variants have a shared mitochondrial dysfunction. Previous work has suggested that *LRRK2* and α-synuclein converge on mechanisms that lead to neuronal death specifically through their effects on the autophagy-lysosomal pathway[174,175]. Given established roles for all three genes in mitophagy, dysfunction of this pathway triggering compensation may be a strong candidate route for producing the rise in expression across monogenic PD cases.

In addition to studying mitochondrial expression profiles in monogenic PD, a parallel aim was to understand how these profiles presented in the sporadic form. Sporadic PD accounts for more than 90% of all PD cases. Several monogenic PD loci have been associated with mitochondrial processes, pointing to mitochondrial dysfunction as integral to PD more generally. However, the mechanisms of mitochondrial dysfunction in sporadic PD have not been fully characterised. To date, no work has been done to characterise mitochondrial expression profiles between sporadic PD cases and controls in a large cohort. As such, the aim here is to fill this gap using whole blood data, meaning that larger sample sizes can be leveraged. In

addition, this study compares these profiles between sporadic and monogenic cases, which may provide evidence towards understanding the extent that mechanisms for mitochondrial dysfunction in monogenic cases can be extrapolated to sporadic PD. Using simple mean comparisons, significant differences between sporadic PD cases and healthy controls could not be reliably observed between the PPMI and PDBP cohorts. Moderately higher expression could be observed in PPMI sporadic PD cases, and no differences could be observed in the PDBP dataset. However, using PCA revealed the existence of subtle hidden structure, namely significant separation of sporadic PD participants and healthy control participants in both cohorts, supporting a role for bioenergetic modulation in sporadic PD.

The second set of transcriptional information analysed in this chapter was mitochondrial PT modification. Significantly decreased modification rates in monogenic PD cases compared to healthy controls were observed at sites 2617 (*MT-RNR2*), 5721 (p9 tRNA, between *MT-CO1* and *MT-ND2*) and 7526 (p9 tRNA, between *MT-CO2* and *MT-CO1*) (Fig. 3.7). The decrease observed in monogenic PD participants is particularly marked at site 2617 (P-adjusted=1.98e-08), which is located within the *MT-RNR2* gene (Fig. 3.5b). Interestingly, the expression increase in genetic cohort PD participants at *MT-RNR2* was also particularly striking (P-adjusted=4.26e-59). Methylation at this site is thought to provide stable interactions to mature mitoribosomes and decreased methylation at this site has been linked to impaired mitochondrial protein synthesis[176]. As such, intuitively one may think that decreased methylation may lead to decreased expression. A potential explanation for the observations here is that reduced protein levels may be driven in part by dysfunctional methylation processes, thus inducing a compensation via increased expression of mtDNA genes. However, further investigation is necessary to test such a model.

One of the key research questions looked at whether mitochondrial transcriptional profiles differed between monogenic PD cases before and after symptom onset. This was aimed at addressing the hypothesis that mitochondrial dysfunction is a core part of disease pathogenesis, rather than a consequence of disease. Di-

mensionality reduction revealed clustering of participants with genetic diagnoses together, regardless of symptomatic status and away from healthy controls (Fig. 3.12). This suggested that the OXPHOS expression profiles of asymptomatic PD participants were more similar to those of symptomatic PD participants than they were to healthy control participants. Recent work leveraged post-mortem brain tissue to study mitochondrial dysfunction in sporadic PD patients[177]. Stratifying across disease severity, the authors reported that mitochondrial dysfunction was detectable before neuronal loss and α-synuclein fibril deposition and state that it is likely that mitochondrial dysfunction is one of the key drivers of early sporadic PD[177]. All in all, the results presented here support the view that mitochondrial dysfunction is likely to be a hallmark of early PD pathogenesis and find that this is detectable in whole blood using expression derived from only 163 OXPHOS genes.

3.4.3 Conclusions and future directions

This chapter addressed two main research questions. The first question aimed to understand whether changes in the mitochondrial transcriptome could be picked up in blood and if so, were changes distinct between monogenic PD cases, sporadic PD cases, and those of healthy control individuals. It was found that mitochondrial transcription profiles (expression and PT modification) of monogenic PD participants were highly distinct compared to healthy controls and that subtle but significant differences between sporadic PD and healthy control participants could be revealed by clustering. The second question aimed to understand whether changes in the mitochondrial transcriptome differed between symptomatic and asymptomatic monogenic PD cases, finding similarities between the mitochondrial expression profiles of symptomatic and asymptomatic cases. In conclusion, this study does observe mitochondrial transcriptional changes in PD cases compared to healthy controls, and these are distinct between monogenic, sporadic and healthy controls, but not distinct between symptomatic and asymptomatic monogenic PD cases. Crucially, these observations were derived from blood samples, and are based on the expression of only 163 genes and modification rates at seven mitochondrial genomic positions. Thus, these findings also contribute to an understanding of mitochondrial transcrip-

tional processes in the disease state across different forms of PD, but further work will be needed to unpick disease causes from consequences.

Cause versus consequence is a challenging concept in ND research, and one which the following chapter attempts to address through application of the *trans*-QTL study paradigm. This method permits a more direct assessment of causality and directionality of effect by exploring whether nuclear genetic variation can explain a proportion of the observed variation in the mitochondrial phenotype. In this way, the following chapter seeks to further understand the role of mitochondrial dysfunction in sporadic PD.

Chapter 4

Understanding mitochondrial transcriptional control in sporadic Parkinson's Disease

4.1 Introduction

Sporadic (or idiopathic) Parkinson's Disease (PD) accounts for 80-90% of PD cases, and yet the underlying disease mechanisms are not well understood, owing to its clinical, pathological and genetic heterogeneity[178]. A number of efforts to understand the biology underlying PD have been made, particularly through genome wide association studies (GWAS) that aim to identify genetic loci associated with sporadic PD. Through these studies, a number of genes, processes and pathways influencing PD risk[65], onset[179] and progression[180,181] have been elucidated. Collectively, these studies and others have generated major insights into the cryptic biology of sporadic PD, identifying roles for a number of pathways with strong evidence supporting the contribution of the endolysosomal, immunological and mitochondrial pathways[182].

The role of mitochondria in PD was identified in the 1980s when chemical interference with mitochondrial complex I gave rise to canonical PD symptoms[183]. Later, genes discovered to be causative for Mendelian forms of PD (e.g. *PRKN*, *PINK1*, *LRRK2*, *DJ1*, *ATP13A2* and *SCNA*) firmly placed mitochondrial pathways

as part of the aetiology of the disease. In sporadic PD, genes involved in mitochondrial pathways have been implicated through GWAS, where common variants within mitochondrial pathways have been found to be associated with sporadic PD status[141], dysregulation of mitochondrial pathways has been observed in sporadic PD monocytes[184] and lymphocytes from individuals with PD were found to have enhanced mitochondrial respiratory activity[185]. As such, mounting evidence points to mitochondrial dysfunction as an important component of sporadic PD aetiology, making this a compelling area for further study.

Despite these important inroads, much remains to be understood. On a granular level, there are many ways in which GWA studies may have missed important genes and biology, for example, many assign gene-to-variant using the 'nearest gene' strategy, which could bias against longer genes[186]. The largest PD GWAS meta-analysis to date identified 90 risk loci including 38 novel independent risk signals, together explaining 22% (range 16-36%) of the heritability of the disease, meaning that a portion of the remaining heritability of the disease is still 'missing' as described by Ohnmacht et al., 2020 and Nalls et al., 2019.

One method commonly used to interpret genotype-phenotype associations detected through GWA studies is to integrate Quantitative Trait Loci (QTL) data, the aim of which is to explain the mechanism underlying the GWAS association. One commonly used QTL is the expression QTL (eQTL). In this case, the variant is thought to be modulating expression of a gene, which could be in *cis* ($<$1Mb away), or in *trans* ($>$5Mb away). In the case of the latter, this could be on the same or a different chromosome. Theoretical work has proposed that $\sim$70% of heritability is driven by small-effect *trans*-acting variants, likely mediated through proximal genes[187,188]. This suggests that identification of these associations has potential for elucidating some of the missing heritabilty of sporadic PD.

The existence of two distinct genomes in the majority of human cells and the necessity for their coordination points to the existence of biologically important *trans*-acting long-range associations. Mitochondria require $\sim$1136 nuclear DNA (nDNA) -encoded proteins to function and previous studies have demonstrated cor-

relation of mitochondrial and nuclear gene expression, supporting an association between the two genomes[75,117,96,3]. Indeed, recent work has identified *trans*-genome acting nuclear variants associated with mitochondrial DNA (mtDNA) -encoded expression and post-transcriptional (PT) modification rates[21,46]. Nevertheless, the contribution of *trans*-genome mitochondrial-nuclear associations to neurodegenerative disease (ND) has not been widely investigated.

The missing heritability of PD, compounded with evidence pointing to a role for mitochondrial dysfunction in its aetiology makes mitochondrial-nuclear coordination an exciting avenue for further understanding sporadic PD. However, no studies have yet looked at the potential role for mitochondrial-nuclear *trans*-acting associations in sporadic PD. In this study, we aim to look at nuclear associations with two mitochondrial transcriptional readouts; mtDNA gene expression and mitochondrial PT modification.

While mtDNA-encoded gene expression is the obvious read out of the mitochondrial transcriptome, mitochondrial PT RNA modification is less commonly examined. PT RNA modifications are changes to nucleic acids in the RNA after transcription. The majority of these are found on transfer RNAs (tRNAs), and they have been confirmed to confer diverse biological functionality[21,10]. In the case of mitochondria, once the mtDNA has undergone polycistronic transcription, modifications are applied to particular nucleic acids of the intermediate and mature RNA. A common PT modification of mitochondrial RNAs (mtRNAs) is methylation (m1A/G) at the 9^{th} position of tRNAs (known as p9 sites) which are known to impact the stability of the secondary structure and affect downstream mtRNA quantities[21,189]. As such, mitochondrial PT modifications are known to have an impact on levels of processed mitochondrial mtRNA, and in this way constitute an important component of mitochondrial transcription. Although limited work has focused on the potential role for mitochondrial PT modifications in disease, a recent study looked at p9 methylation in subjects diagnosed with NDs, where the authors report hyper-methylation at p9 sites detected in cerebellar tissue from subjects diagnosed with Alzheimer's disease (AD) and progressive supranuclear palsy (PSP)[20]. This

is encouraging, and supports the utility for understanding the role of p9 methylation in other NDs. Insofar as I am aware, however, mitochondrial PT modification has not been studied in the context of PD.

As such, the central aim of this study was to interrogate nuclear genetic modulation of mitochondrial transcriptional phenotypes and to discern whether differences could be observed between sporadic PD and healthy control participants. This may offer insights into sporadic PD aetiology and the involvement of mitochondria. To this end, mitochondrial-nuclear *trans*-eQTLs and *trans*-meQTLs (methylation QTLs), which will be referred to collectively as *trans*-xQTLs, were calculated from transcriptomic data for 784 sporadic PD and 507 healthy control participants, derived from the multi-cohort whole blood AMP-PD dataset. Additionally, this study aimed to focus on variant-to-gene assignation, utilising newly available high fidelity HiC chromatin capture data with the aim of robust candidate gene identification, aiming to identify novel biology highly relevant to sporadic PD.

4.2 Methods

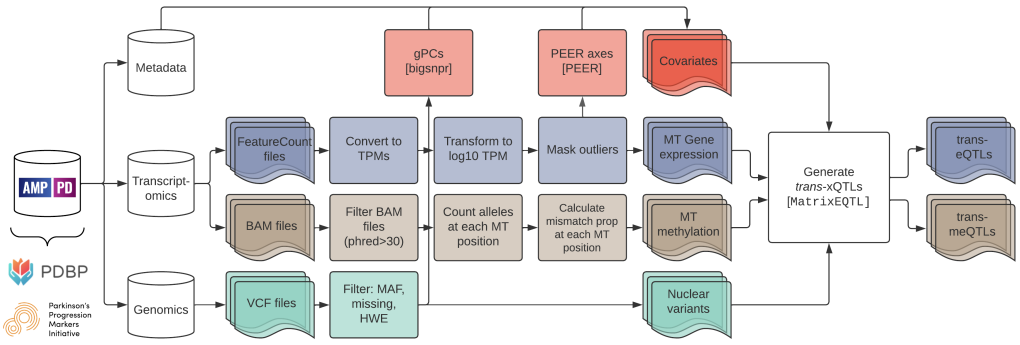

Figure 4.1: Workflow diagram showing the analysis pipeline followed in chapter 3.

4.2.1 Data

Data from the Accelerating Medicines Partnership (AMP) for Parkinson's (PD) - derived cohorts, Parkinson's Progression Markers Initiative (PPMI) and Parkinson's Disease Biomarkers Program (PDBP) cohorts were obtained by access request as described in section 3.2.1. The analyses in this chapter, as illustrated in figure 4.1

leveraged transcriptomic, genomic and metadata from the PPMI and PDBP cohorts. Other than cohort-specific sample selection protocols, an identical pipeline was applied across cohorts (Fig. 4.1).

4.2.1.1 PBDP participant selection strategy

To maximise participant numbers, PDBP was used as the discovery dataset for the analyses in this chapter as it comprised the largest number of sporadic PD diagnosed participants of all AMP-PD cohorts. From this cohort, to obtain a static dataset containing unique participants only, "visit_month" 0 was selected for the *trans*-xQTL analysis as the largest number of participants were measured at this timepoint (see Fig. 3.2c for cohort structure details). Non-standard controls as well as some case participants with ambiguous assignations were present in the dataset, and as such sporadic PD cases and healthy controls were selected from the M0 data using the "study_arm" variable available in the clinical metadata, where sporadic PD cases were indicated with "PD" and healthy controls with "Healthy Control". In the same vein, 12 participants were filtered out as they had "case_control_other_latest" recorded as "Other" which represented an ambiguous diagnosis. Additionally, participants with conflicting assignations were removed, which included those recorded with "study_arm" as "Healthy Control" and "case_control_other_latest" as "Case" and vice-versa. This left a total participant number of 507 in the healthy control group and 784 in the sporadic PD group.

4.2.1.2 Genetic data

Blood samples collected at clinics on a cohort-specific basis were the starting point for the harmonised sequencing and processing workflow implemented by AMP-PD. DNA extracted from participant whole blood specimens was sequenced using the Illumina HiSeq X Ten platform[190]. Paired-end 300-400bp reads were processed according to the harmonisation pipeline outlined by Regier et al., (2018) at the Broad Institute[191]. Alignment and variant calling were performed against reference genome GRCh38[190]. Per-chromosome variant call files (VCFs) were downloaded from the AMP-PD Google Cloud bucket to a local server using the gsutil tool. These were then split into per-chromosome, per-participant files before

merging to per-participant files using VCFtools (v0.1.16). To ensure that only high quality, common variants were analysed, the VCF files were filtered using vcftools for minor allele frequency (MAF) >1%, missing data <1% and a Hardy-Weinberg Equilibrium P-value of >0.001[192]. This resulted in a total of 5,815,014 variants carried forward for downstream analyses.

4.2.1.3 Mitochondrial phenotype data

Both mtDNA gene expression data and PT modification data underwent near identical processing as outlined in section 3.2.2.1 and section 3.2.2.2. However, crucially, correction was not directly applied to the data, instead, covariate files were supplied to MatrixEQTL for generation of *trans*-xQTLs. Additionally, the modification rate data was not transformed in any way prior to running MatrixEQTL, as was done for the distribution-type analyses in chapter 3.

4.2.1.4 Metadata

Sample and participant -level metadata were aggregated and downloaded as per section 3.2.2.3.

4.2.1.5 Covariates

Covariates were supplied to MatrixEQTL and were comprised of key metadata variables, genetic principle components (gPCs) and axes of variation generated using probabilistic estimation of expression residuals (PEER factors). From the metadata, age and sex were included in the covariate file supplied to MatrixEQTL, as discussed in section 3.2.2.1. PEER has previously been shown to improve eQTL discovery power in bulk RNA-seq data, and is widely used for these analyses[108]. As such, PEER was used here to capture hidden confounders such as cell type composition and batch effects. The standalone command line PEER tool (version 1.3) was run on the TPM matrices with 1000 iterations. The first 10 PEER factors were included in the MatrixEQTL covariate file as per Ali et al., 2020. The same 10 PEER factors were utilised for mapping both eQTLs and meQTLs. To capture population substructure, gPCs were calculated using the bigsnpr R package[193]. The input file for this was a browser extensible data (BED) format file generated from

the VCF file using plink2. Missing values are not tolerated by the "snp_autoSVD" function and as such, imputation was performed using the "snp_fastImputeSimple" function, setting the "method" argument to "random". The "snp_autoSVD" function was then run setting thr.r as 0.2 (this is the default recommended value and allows MAF-based SNP clumping which removes long-range LD regions) and k as 10 (number of PCs to calculate). The first 5 gPCs (as per Ali et al., 2020) were included in the MatrixEQTL covariate file.

4.2.2 xQTL mapping

To generate xQTLs, the MatrixEQTL R package was used which takes phenotype, genotype and covariate matrices (generation described above) as input[194]. MatrixEQTL was run in two modes, the first using a simple linear regression model was applied to sporadic PD case data and to healthy control data separately. This was achieved by setting the "useModel" argument in the main function "Matrix_eQTL_engine" to "modelLINEAR". The formula for "modelLINEAR" was as follows:

$$phenotype = \alpha + \sum_{k} \beta_k \cdot covariate_k + \gamma \cdot genotype\_additive \tag{4.1}$$

The second mode used a special case of a linear regression model and was applied to combined sporadic PD case and healthy control data files in order to test for the significance of the interaction between the genotype and the last covariate. Associations generated using this model are referred to in this chapter as interaction term xQTLs, or DxG xQTLs. As such, the input files were similar to those required by the first mode, but here all participants were included in the same input files and the covariate file had one additional variable, which was a dummy variable indicating the grouping (i.e. a 1 for sporadic PD cases, and a 0 for healthy controls). This model was applied by setting the "useModel" argument in the main function "Matrix_eQTL_engine" to "modelLINEAR_CROSS". The formula for "modelLINEAR_CROSS" was as follows:

$$phenotype = \alpha + \sum_k \beta_k \cdot covariate_k + \gamma \cdot genotype\_additive + \delta \cdot genotype\_additive \cdot covariate_k \quad (4.2)$$

Other arguments supplied to "Matrix_eQTL_engine" were "noFDRsaveMemory" set to "FALSE", so that a false discovery rate (FDR) adjusted P-value was included in the output files. Additionally, "pvalue.hist" was set to "qqplot" to output a qqplot for each run, allowing quality control checking of the output data.

MatrixEQTL was run in a per-phenotype per-model fashion, generating an xQTL file for each phenotype-model combination. For example, to generate eQTLs for phenotype MT-ND6 gene expression, a MatrixEQTL run was performed for sporadic PD cases (under model 4.1), for healthy controls (under model 4.1) and including all participants and a grouping variable to obtain the interaction effect (under model 4.2).

Lead (peak) SNPs were defined as those with the lowest P-value in a 1 megabase (Mb) block. These were determined using a custom function. First, the "hclust" function from the base R "stats" package clustered all variant positions within a chromosome, generating a distance tree representing the distance between all variant pairs. The distance tree was then cut using the stats "cutree" function into clusters each spanning 1Mb. An ID could then be assigned to each xQTL association representing the chromosome and Mb block it belonged to. All xQTL associations were then grouped by their ID, and the association with the minimum P-value in the group was selected as the peak SNP.

The final hits given in table 4.1 were determined by using an FDR cut-off of 0.05 for those observed under the standard linear model (4.1) and a less stringent cut-off of 0.1 for those observed under the DxG model (4.2), as disease-associated *trans*-xQTLs are challenging to observe.

4.2.3 xQTL replication

Significant associations were replicated in a second AMP-PD cohort, PPMI. The M36 timepoint was selected, the logic for which is outlined in section 4.2.1.1 and

an identical pipeline was run as for PDBP timepoint 0 (PDBP-M0). Following this, associations identified in healthy individuals from PDBP were replicated in independent healthy control datasets derived from Ali et al., 2020 for eQTLs and from Ali et al., 2019 for meQTLs. In these studies, human genome build 37 was used, whereas AMP-PD used build 38. As such, genomic co-ordinate lift-over was performed using the "convert_rs_to_loc" function from the colochelpR package[195]. For all replication, the identical variant or the closest variant in high LD with the identical variant was considered as representative of the signal. To get variants in LD with a target variant, a bespoke, and now publicly available software package was developed, ensemblQueryR. The "ensemblQueryLDwithSNPlist" function was used, setting both the minimum R2 and D' to 0.8. This R package can be found at the following URL: github.com/ainefairbrother/ensemblQueryR.

4.2.4 Downstream analyses

4.2.4.1 Annotation

To annotate the xQTLs with a candidate causal gene (nearest gene, V2G-nominated gene), most severe consequence and other relevant information, a custom wrapper was written utilising the Open Targets API to automate the process of pulling down Open Targets information on a variant ID basis. This code was made publicly available and can be accessed at the following URL: github.com/ainefairbrother/useful-bioinformatics-functions/blob/main/helper-funs-to-query-api-opentargets-gh.R.

4.2.4.2 Mediation analysis

Mediation analysis was run with the purpose of understanding whether, for a given xQTL, the variant was acting by mediating the expression of a gene in *cis* to modulate the mitochondrial phenotype (Fig. 4.2 shows a schematic of this model). This was done using custom code which leveraged the latent variable modelling R package, Lavaan[196]. The mediation analysis model included three variables: the nuclear variant, the *cis* nuclear gene and the *trans* mitochondrial phenotype. Accordingly, the model was set up to include 3 paths: path "A" was the effect of the variant on the *cis* gene, path "B" the effect of the *cis* gene on the *trans* phenotype and path

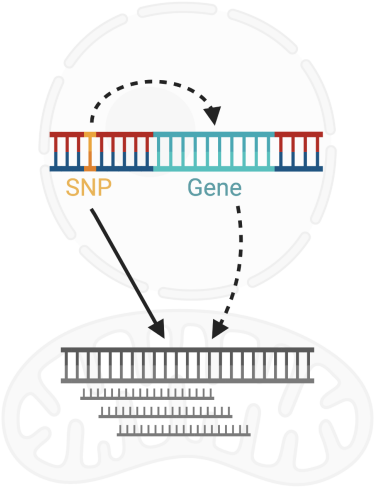

Figure 4.2: Schematic to show the scenario tested using mediation analysis. In this example, a nuclear SNP is modulating mtDNA gene expression in *trans* through a nuclear gene in *cis*.

"C" the direct effect of the variant on the *trans* gene. A further model term was included to capture the indirect effect of the variant on the *trans* gene via the *cis* gene. This model was defined in the Lavaan syntax as follows:

```
"
# Path C
trans_gene ~ c*variant

# Path A
cis_gene ~ a*variant

# Path B
trans_gene ~ b*cis_gene

ab := a*b
"
```

This model was then fitted to the underlying data for each target xQTL. For example, for a given eQTL, this would comprise a data frame with three columns: 1.

genotype, 2. *cis* gene expression and 3. mtDNA gene expression (*trans* phenotype). The model fit was then performed on this data using the Lavaan "sem" function, setting the "se" argument to "bootstrap", which selects the bootstrap function and the "bootstrap" argument to 5000, which represents a reasonable balance between speed and robustness. The bootstrap method draws random samples from the data, running the analysis on sequential subsets. This is particularly useful for input data that is not normally distributed. Once the model fit was performed, a second Lavaan function, "parameterEstimates", was used to estimate the parameters of the model, generating P-values and test statistics for each path included in the model. To assess whether mediation was likely to explain an xQTL, the P-value and test statistic corresponding to the "ab" model parameter were extracted. Statistical significance for mediation was then assessed at two levels, nominal ($P<0.05$) and multiple test corrected Bonferroni ($P<(0.05/\text{no. of tests})$).

4.2.4.3 Chromatin contact analysis

To improve the accuracy of candidate casual gene assignments, HiC chromatin contact data was levaraged. HiC can indicate whether a variant is in a region of high chromatin contact for the *cis* candidate gene, as such, even if a variant is linearly distal, in the true 3D structure of the genome, the variant may be closer, and as such may be able to modulate the gene in *cis*. In light of this, data from a recent study that generated high-density chromatin contact maps was leveraged[197]. Hi-C data was downloaded on a per-gene basis from the following repository: https://gillisweb.cshl.edu/HiC/. The data was then plotted using a custom R script, and relevant variant and *cis* gene positions were overlaid for interpretation of the target xQTLs.

4.3 Results

Genes linked to mitochondrial processes have been identified to be causal for monogenic familial PD, and have been implicated in sporadic PD. However, the extent to which this is causal and which specific processes might be involved are not currently understood. To consider this we study two functional readouts of the mitochondrial

genome (expression and PT modification) and test whether the genetic regulation of these processes by the nuclear genome is different in sporadic PD patients versus healthy controls.

4.3.1 Discovery of mitochondrial-nuclear *trans*-xQTLs

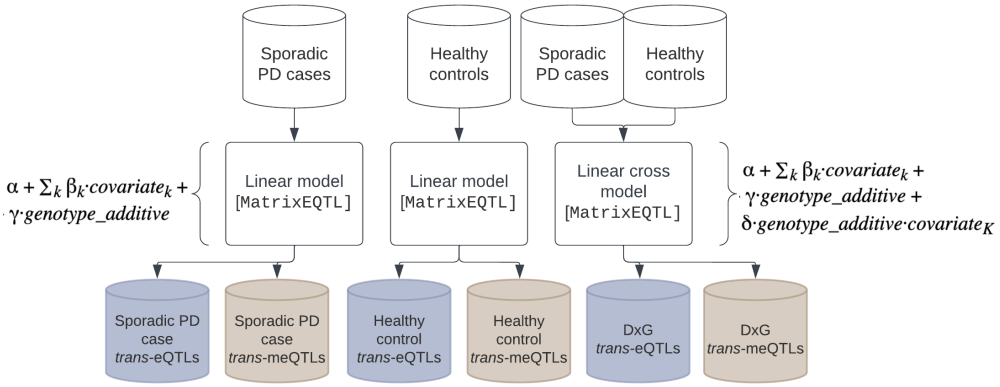

Figure 4.3: Diagram to show the experimental design for the *trans*-xQTL analysis. Formulae show the two models (adapted from Shabalin et al., 2021 [194]), implemented using MatrixEQTL, that were utilised for the analyses in this study, as well as the the input and output data.

In order to identify expression Quantitative Trait Loci (eQTLs) and methylation Quantitative Trait Loci (meQTLs) - collectively termed xQTLs - the PDBP cohort (at timepoint "M0", see section 4.2.1.1) was analysed using MatrixEQTL. A simple additive linear model was run separately on cases (n=784) and controls (n=507) (Fig. 4.3). The same model was run on cases and controls together including a genotype-by-disease (G×D) interaction term to detect sporadic PD -specific effects. A significant G×D interaction term association between a SNP-gene pair indicates that the effect of genotype on expression is significantly different in sporadic PD compared to healthy controls. Biologically, this formal analysis framework can therefore detect: 1. xQTLs significant in healthy control participants but not significant in sporadic PD participants, 2. xQTLs significant in sporadic PD participants but not in healthy control participants, 3. xQTLs with opposing directions of effect in healthy control participants compared to PD, and 4. xQTLs whose effects in healthy control participants and sporadic PD participants have the same di-

rectionality but have significantly different magnitudes. The interaction term in the genotype-expression regression is used to determine whether there is a statistically significant difference in the slope of the regression line between healthy control individuals and individuals with sporadic PD. Specifically, this term assesses whether the effect size of a unit change in allele dose on the phenotype is significantly different between sporadic PD participants and healthy control participants.

Therefore, this study seeks to identify nuclear genetic regulation of mitochondrial processes in sporadic PD cases, healthy controls and those specific to sporadic PD disease (G×D) for two mitochondrial transcriptional phenotypes: mtDNA gene expression (15 mtDNA-encoded genes) and rate of PT modification (methylation) of mtDNA-encoded RNA (at 6 functionally important positions). After performing these analyses, in total for expression from mtDNA-encoded genes, 8 G×D interaction term associations were detected (FDR 10%, table 4.1). Furthermore, two sporadic PD associations and one healthy control association were detected (FDR 5%, table 4.1). For modification at mitochondrial genomic positions, 18 G×D interaction term associations (FDR 10%, table 4.1) were detected. Additionally, no sporadic PD associations and 10 healthy control associations were detected (FDR 5%, table 4.1). In all cases, these associations represent significant interactions between peak nuclear encoded genetic variants, and unique mitochondrial RNA phenotypes (or interaction terms).

Model	Mito phenotype (gene/position)	rsID	CHR	BP	A1	BETA	P	SNP type	Nearest gene TSS	OpenTargets V2G nominated gene	Mediating gene
Interaction (DxG)	MTND6	rs7130955	11	65605109	T	-	2.692581e-09	intron_variant	KCNK7	RNASEH2C	RNASEH2C, MUS81, CTSW, MAP3K11, FAU
Interaction (DxG)	MTRNR1	rs851811	1	1.18E+08	C	-	2.422406e-08	intergenic_variant	TENT5C	TENT5C	CD101
Interaction (DxG)	MTCO3	rs11491186	17	11973393	A	-	5.295960e-08	intron_variant	ZNF18	DNAH9	
Interaction (DxG)	MTND6	rs695550	22	25836560	A	-	1.359430e-07	intron_variant	MYO18B	MYO18B	
Interaction (DxG)	MTND6	rs6429398	1	2.43E+08	C	-	3.670320e-07	intergenic_variant	CEP170	CEP170	
Interaction (DxG)	MTND6	rs146832598	11	8274057	A	-	7.134903e-07	intergenic_variant	LMO1	LMO1	
Interaction (DxG)	MTND6	rs77809042	8	72211805	C	-	9.690770e-07	intron_variant	TRPA1	MSC	
Interaction (DxG)	MTND6	rs4976988	8	1.42E+08	C	-	9.823654e-07	downstream_gene_variant	TSNARE1	ADGRB1	
Case	MTND4L	rs9868989	3	2768413	A	8.664739e-02	2.982572e-09	intron_variant	IL5RA	CNTN4	
Case	MTCO1	rs4907234	2	96447202	G	-4.175226e-02	1.816113e-08	intergenic_variant	NEURL3	GPAT2	ITPRIPL1, NCAPH
Control	MTCO3	rs4724362	7	45121704	C	-7.406997e-02	1.669464e-08	intergenic_variant	TBRG4	CCM2	MYO1G, CCM2, ZMIZ2, OGDH, NUDCD3
Interaction (DxG)	5883	rs7320769	13	22935164	A	-	1.633474e-08	intergenic_variant	SGCG	SGCG	
Interaction (DxG)	5883	rs73044588	12	6252835	T	-	1.772833e-08	intergenic_variant	CD9	CD9	
Interaction (DxG)	5883	rs75136605	19	2637251	A	-	9.186513e-08	intron_variant	GNG7	GNG7	
Interaction (DxG)	5883	rs9869896	3	65616723	C	-	9.459236e-08	intron_variant	MAGI1	MAGI1	
Interaction (DxG)	5883	rs4700264	5	58237197	T	-	2.115414e-07	intergenic_variant	PLK2	PLK2	
Interaction (DxG)	5883	rs9892440	17	61287458	T	-	2.522277e-07	intron_variant	TBX2	TBX2	
Interaction (DxG)	5883	rs72966766	2	2.22E+08	C	-	2.527593e-07	intron_variant	SGPP2	SGPP2	
Interaction (DxG)	5883	rs7431895	3	85821351	C	-	2.835633e-07	intron_variant	CADM2	CADM2	
Interaction (DxG)	5883	rs10267108	7	15101717	G	-	3.088262e-07	intergenic_variant	DGKB	DGKB	
Interaction (DxG)	5883	rs10159481	10	37362424	C	-	3.704534e-07	intron_variant	ANKRD30A	ZNF25	
Interaction (DxG)	5883	rs58883854	14	94927064	T	-	4.287767e-07	intergenic_variant	GSC	GSC	
Interaction (DxG)	5883	rs7702630	5	67609863	G	-	5.322697e-07	intron_variant	CD180	PIK3R1	
Interaction (DxG)	5883	rs10516330	4	18801778	C	-	6.859870e-07	intergenic_variant	LCORL	-	
Interaction (DxG)	5883	rs140918712	11	77138321	CGCCGACCCT	-	7.659697e-07	intron_variant	MYO7A	MYO7A	
Interaction (DxG)	5883	rs17119278	10	1.06E+08	T	-	7.894911e-07	intergenic_variant	SORCS3	-	
Interaction (DxG)	5883	rs867397752	4	1.62E+08	A	-	8.682974e-07	intron_variant	FSTL5	FSTL5	
Interaction (DxG)	5883	rs6092348	20	56667920	G	-	9.691643e-07	downstream_gene_variant	TFAP2C	TFAP2C	
Interaction (DxG)	5883	rs33947506	4	16884933	GA	-	1.144820e-06	intron_variant	LDB2	LDB2	
Control	7526	rs12602609	17	3294002	C	1.413777e-02	1.255536e-09	intron_variant	OR3A1	OR1E1	OR3A2, SPATA22, OR1E1
Control	5883	rs10161868	13	55636511	A	3.398980e-04	1.367182e-09	intergenic_variant	PRR20C	-	
Control	5883	rs9860933	3	65609693	C	3.311355e-04	1.583054e-08	intron_variant	MAGI1	MAGI1	
Control	5883	rs548992	9	1.32E+08	T	-2.863891e-04	1.643688e-08	intergenic_variant	SETX	SETX	
Control	7526	rs12324408	15	39757111	T	1.558199e-02	5.086937e-08	intron_variant	FSIP1	THBS1	BUB1B, PAK6
Control	5883	rs111949974	5	1.74E+08	T	2.615813e-04	6.249215e-08	intron_variant	NSG2	NSG2	
Control	5883	rs569426832	2	1.92E+08	AAG	2.747839e-04	7.610604e-08	intergenic_variant	NABP1	MYO1B	
Control	7526	rs35598857	1	38055579	A	1.338011e-02	4.520793e-07	intron_variant	POU3F1	FHL3	
Control	7526	rs17207766	17	9808491	T	1.312140e-02	5.396811e-07	intron_variant	GSG1L2	GSG1L2	
Control	7526	rs2673311	6	1.12E+08	C	-7.211182e-03	6.161649e-07	intergenic_variant	CCN6	FAM229B	

Table 4.1: Table to show all xQTL associations under FDR<0.05 for healthy control and sporadic PD case associations, and under FDR<0.1 for (G×D) interaction term associations.

4.3.2 Replication of *trans*-xQTLs

All peak associations passing FDR<0.05 under the standard linear model or FDR<0.1 under the G×D model detected in PDBP (as listed in table 4.1) were considered for replication in an independent data set. Replication was performed first in a pair of entirely independent healthy control eQTL and meQTL datasets derived respectively from Ali et al., (2019)[46] and Ali et al., (2020)[21], which although permits replication of healthy control xQTLs only, provides independence of sample collection, processing, sequencing and downstream analysis. Following this, replication was carried out in a second AMP-PD cohort (PPMI, timepoint "M36"), allowing replication of all xQTL categories (sporadic PD case, healthy control and G×D interaction term associations).

First, replication was performed using independent xQTL data. As no datasets containing PD nuclear-mitochondria *trans*-eQTLs or *trans*-meQTLs have previously been generated, to my knowledge, healthy control participant data sets were utilised for replication, limiting the scope of replication to healthy control xQTLs only. Two datasets, one containing eQTLs, one containing meQTLs are meta-analyses which each aggregate associations from from 4 separate data sets. First looking at eQTL replication, the single healthy control association replicated strongly ($P_{replication}$=6.71e-18). Next, looking at meQTL replication, the present study detects healthy control associations with modification rates at two different mitochondrial genomic sites, 5883 and 7526. However, the replication meQTL dataset only includes associations for 7526, so only healthy control signals from this site can be considered for replication. Under a Bonferroni adjusted P-value correcting for the total number of peak associations being tested (P<0.05/6, where 6 is the overlap between healthy control xQTL phenotype associations detected and phenotypes available in the replication data), 0/5 healthy control associations replicated. Equally, using a nominal P-value cut-off (P<0.05), 0/5 healthy control associations replicated.

Next, replication was performed in the PPMI dataset, for which xQTLs were generated using the same pipeline as for the PDBP (discovery) xQTLs. First look-

ing at eQTL replication, 11/11 associations were also generated in the PPMI data. Under a Bonferroni adjusted P-value correcting for the total number of peak associations being tested (P<0.05/11), 0/8 G×D interaction term associations, 1/1 healthy control associations and 0/2 sporadic PD case associations replicated in PPMI. Using a nominal P-value cut-off (P<0.05), 2/8 G×D interaction term associations, 1/1 healthy control associations and 0/2 sporadic PD case associations replicated in PPMI. Next, looking at meQTL replication, 21/28 of associations were also detected in the PPMI data. Under a Bonferroni adjusted P-value correcting for the total number of peak associations being tested (P<0.05/21), 0/21 G×D interaction term associations and 1/7 healthy control associations replicated in PPMI. No sporadic PD case associations were detected in PDBP, and as such there were none to replicate. Using a nominal P-value cut-off (P<0.05), 1/21 G×D interaction term associations and 3/7 healthy control associations replicated in PPMI.

Considering that the G×D interaction term associations are each comprised of a healthy control association and a sporadic PD case association with significantly different beta slopes, to further interrogate the replicability of the G×D associations, their components, or "arms" were replicated. For each G×D association, replication was performed on the case and control arms if they reached a nominal P-value (P<0.05). First, looking at eQTL G×D replication, considering the 8 G×D associations, 5 had case and control passing P<0.05 and 1 had control passing P<0.05 and as such, 6 were considered for replication. Of the 5 with case and control passing P<0.05 in PDBP, 2/5 had cases replicating in PPMI with one passing Bonferroni significance (P<0.05/6) and the other nominally significant (P<0.05). For the one association with only control passing P<0.05 in PDBP, this did not pass P<0.05 in PPMI. As such, 2/6 (33%) of PDBP G×D expression associations had at least one of their association arms replicating in PPMI (Fig. 4.4). Next, looking at meQTL G×D replication, considering the 18 G×D associations, 8 had case and control passing P<0.05 and 10 had one of case or control passing P<0.05. As such, all 18 were considered for replication. 14/18 associations were detected in the replication data, of which 3/14 (21%) of PDBP G×D modification rate associations

had at least one of their association arms replicating in PPMI (Fig. 4.4). Here, it is important to note that G×D associations are likely to be difficult to replicate in PPMI due to the sample sizes available.

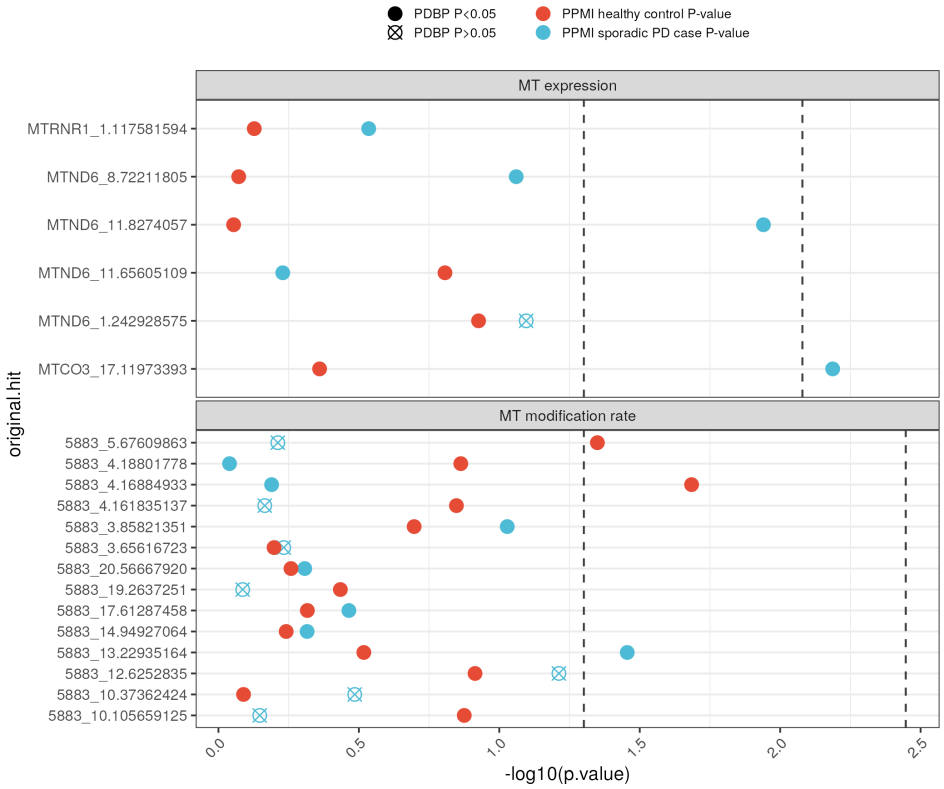

Figure 4.4: Figure to show replication (in PDBP) of PPMI-discovered case and control eQTL and meQTL associations that comprise each significant G×D association. The top facet shows eQTLs, with the first dashed line representing P=0.05 and the second representing Bonferroni corrected P=0.05/6. The top facet shows meQTLs, with the first dashed line representing P=0.05 and the second representing Bonferroni corrected P=0.05/14. Filled points indicate that the association had a P-value of less than 0.05 (nominally significant) in the replication dataset (PDBP), whereas empty points with a cross indicate that the association had a P-value greater than 0.05 (non-signigficant) in the replication dataset (PDBP). Red points are the healthy control arm of the G×D association, whereas blue points are the sporadic PD arm of the G×D association.

Finally, to check whether top associations derived from an independent study replicate in the present study, the 18 top "whole blood" associations from the same independent healthy control eQTL study discussed above were checked for their significance levels in the PDBP eQTLs. 6/18 (33%) of these associations were detected at P<0.05, 5/6 of which are an association of the expression of 5 mtDNA

genes with variant rs6973982 (1e-03<P<1.7e-08), and 1/6 is an association between MT-RNR2 and rs10165864 (3.6e-02).

It is important to note two factors that make replication challenging in this instance. Firstly, the sample sizes of the PPMI replication dataset are considerably smaller than the discovery dataset (135 healthy controls and 302 sporadic PD cases) making it difficult to replicate any but the largest signals. Secondly, with respect to the RNA-Seq data used to generated the independent replication meQTL dataset, an RT enzyme with lower fidelity was utilised during library preparation meaning that PT modification rates could be captured more readily in these data. As a consequence, replication of PDBP-generated meQTLs in particular presented significant challenges.

All considered, strong replication of healthy control eQTLs could be observed in a second AMP-PD cohort and in independent meta-analyses. Independently derived healthy control eQTLs were also detected as significant associations in the PDBP-generated eQTLs. G×D interaction term associations were more replicable than sporadic PD case associations and this was supported by replicability present within the comprising cases and controls, even if the overall G×D P-value could not be replicated. Generally, eQTLs were more replicable than meQTLs, which were considerably more challenging to replicate.

4.3.3 Analysis and interpretation of *trans*-xQTLs

4.3.3.1 Interaction term sporadic PD -specific *trans*-xQTLs

To identify xQTLs that are specific for sporadic PD in a statistically robust manner, the PDBP data was analysed using a linear model with a G×D term. A significant G×D interaction term indicated that the effect of the genotype on the mitochondrial phenotype was significantly different in PD participants compared to healthy controls.

The most significant interaction term eQTL identified under the G×D model was the association between *MT-ND6* and rs7130955 (11:65605109) (P=2.69e-09), characterised by a positive association in healthy control participants (P=8.9e-03, beta=0.064) and a negative association in sporadic PD participants (P=6.97e-03,

beta=-0.047) (Fig. 4.5a and b). Supporting its robustness, this association was also detected in the PPMI dataset at nominal significance (P=0.0421). To further understand the potential genes and mechanisms through which rs7130955 is associated with expression of *MT-ND6*, a number of evidence streams were considered. Firstly, it was important to note that rs7130955 is located in the intronic region of the *MAP3K11* gene. Whilst this contributes to evidence for *MAP3K11* as the candidate gene for *cis* regulation by rs7130955, it does not guarantee it. Where it is often the case that variants regulate proximal or host genes, it is also known that variants can regulate distal genes instead[198]. As such, further interrogation of gene candidacy for rs7130955 was performed.

In the first instance, the Open Targets Genetics variant-to-gene (V2G) algorithm was leveraged. This approach seeks to assign variants to genes by integrating available evidence to generate a V2G score. Dependent on data availability for the target variant, evidence may include the linear distance between the SNP and transcription start site (TSS), molecular phenotype experiment data (eQTL, splicing QTL (sQTL), protein QTL (pQTL)), chromatin interaction experiment data and functional prediction (Ensembl VEP)[199]. The highest scoring V2G gene for rs7130955 was the ribonuclease subunit coding gene *RNASEH2C* (V2G=0.32), evidenced by linear distance to the TSS (115,709bp), sQTL data (sQTL in 19/49 GTEx tissues, excluding whole blood) and eQTL data (eQTL in 11/126 studies, including GTEx whole blood). Five other genes with V2G>0.25 for rs7130955 included the novel Rab8-binding protein gene *EHBP1L1* (V2G=0.29), the transcriptional repressor gene *OVOL1* (V2G=0.29), the mitogen-activated protein kinase gene *MAP3K11* (V2G=0.28), the lysine acetyltransferase gene *KAT5* (V2G=0.26) and the nuclear assembly factor *BANF1* (V2G=0.26). It is important to consider that the V2G score integrates signals arising from diverse tissues as opposed to purely whole blood which is the focal tissue of this study. Considering the V2G scores and whole blood eGenes, four likely candidates were taken forward: *MAP3K11*, *RNASEH2C*, *KAT5* and *BANF1*. To understand which genes have previously been identified as eGenes for rs7130955 in whole blood, the second stream of evidence considered only whole

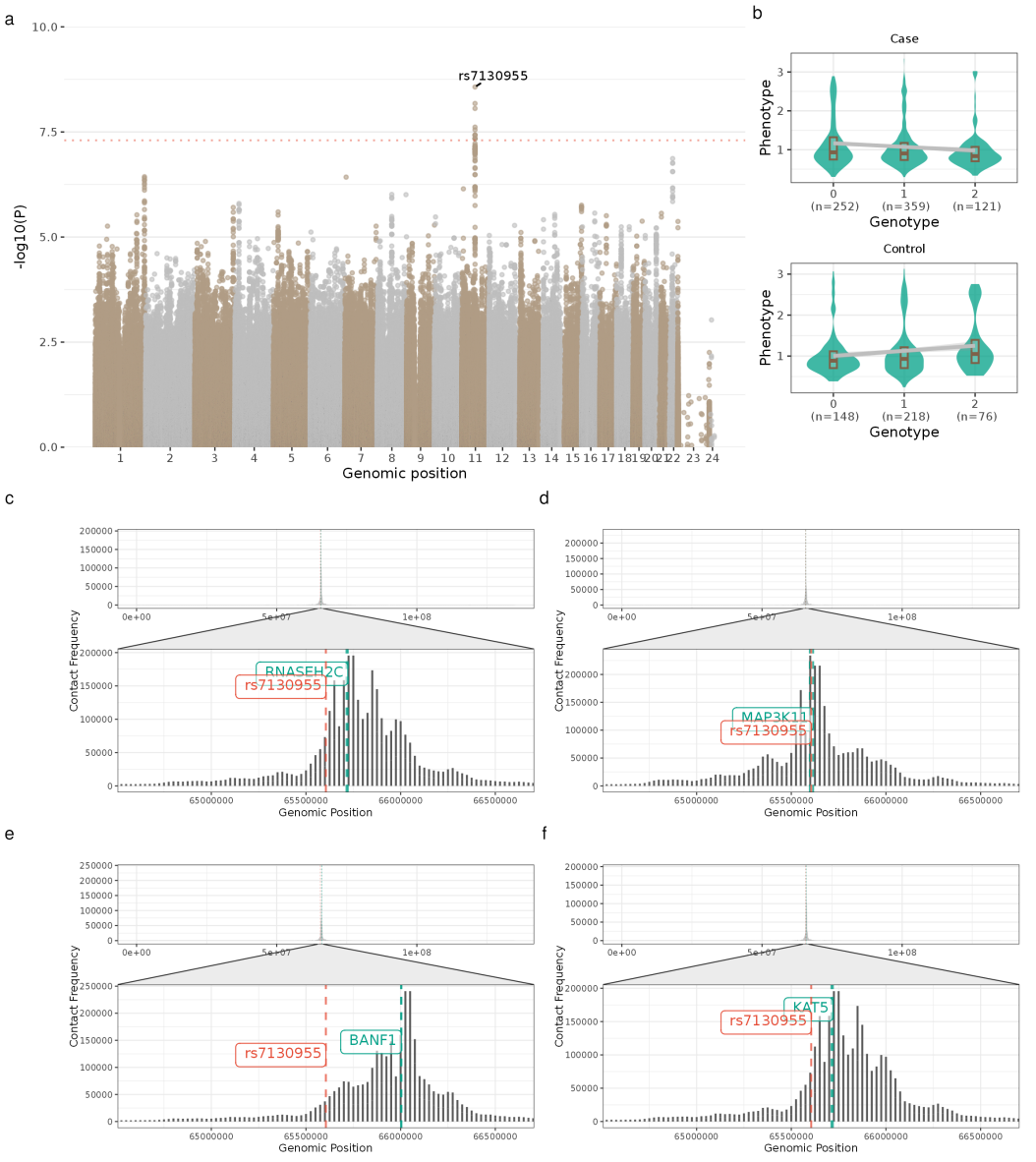

Figure 4.5: A. Manhattan plot to show all G×D associations between nuclear variants and expression of the mtDNA gene MT-ND6. B. Genotype-by-phenotype distributions, where each data point represents a study participant, underlying the association between rs7130955 and MT-ND6 expression in sporadic PD cases (top) and healthy controls (bottom). C-D. Plots to show Hi-C chromatin contact data between a chromosome and its host gene (at a resolution of 25kbp), where the gene location is indicated by two vertical green lines (start, end) and a corresponding green label. The position of variant rs7130955 is indicated by a red vertical line and corresponding red label. The zoomed area represents a 1Mbp window around the variant of interest.

blood *cis*-eQTL data. The eQTLGen data is a meta-analysis of 37 whole blood eQTL datasets, considering 11 million SNPs and identifying 16K eGenes[200]. In the eQTLGen data, the top eGene for rs7130955 was *MAP3K11* (P=3.27e-310). *BANF1* (P=8.8833e-42), *RNASEH2C* (P=5.34e-39) and *KAT5* (P=1.16e-8) were also significant eGenes. Notably, the V2G score for *MAP3K11* integrates across 4 blood data sets (GTEx, Lepik 2017, TwinsUK and eQTLGen), and in 3/4, *MAP3K11* was identified as an eGene for rs7130955 (beta>0), whereas only 1/4 data sets (eQTLGen) identified *BANF1* and *RNASEH2* as eGenes for rs7130955.

A common mechanism by which *trans*-eQTLs operate is that a variant regulates a gene in *cis* (in this case, a nuclear gene), which in turn modulates a phenotype in *trans* (in this case, mitochondrial expression). As such, mediation analysis was carried out, testing the model that rs7130955 is acting through genes in *cis* to modulate the mitochondrial phenotype. Mediation analysis was run on all genes with a TSS less than 500kb from rs7130955 (47 genes in total). Only *RNASEH2C* (P=1.97e-04) was significant at a Bonferroni adjusted P-value of P<0.05/47. However, evidence for mediation was found at a nominal P-value of P<0.05 for four other genes, including the eQTL- and V2G- nominated *MAP3K11* (P=0.018).

The final stream of evidence considered to identify the candidate gene for rs7130955 was that of chromatin contact data. When trying to understand gene regulation, it is important to consider the 3D structure of the genome and its role in gene regulation. In the context of eQTLs, a recent study using chromatin conformation capture (Hi-C) found close spatial proximity between eQTLs and their eGenes across multiple human tissues[201]. In light of this, further interrogation of the rs7130955-*MT-ND6* eQTL aimed to understand which gene rs7130955 is closest to in the 3D structure of the genome, and as such, which gene is more likely to be the eGene. Whilst the V2G gene dataset does utilise Hi-C data from the Javierre et al., (2016)[202] study, the present analysis leverages the more recent deep, robust and high-powered Hi-C Lohia et al., (2022) meta-analysis[197]. Lohia et al., (2022) generated a 'global high-density chromatin interaction network' by meta-analysing 3619 human Hi-C runs. For each candidate eGene, chromosome-wide contact fre-

quencies (contact of the loci with every other genomic loci on the host chromosome) for each were plotted, overlaying the candidate eGene bounds (green) and the variant position (red) (Fig. 4.5). In figure 4.5d, the chromosome 11 -*MAP3K11* contact map is shown, with the zoom focusing on the locus of interest. rs7130955 is within the highest contact frequency bin (>200,000) which is not surprising given the small linear distance between them. Comparing this to the other three candidate eGenes, rs7130955 falls into the 75,000 contact frequency bin for *RNASEH2C* and *KAT5*, and into the <50,000 bin for *BANF1*. *RNASEH2C*, *KAT5* and *BANF1* thus display significant but considerably less frequent physical contact with rs7130955 than observed for *MAP3K11* (Fig. 4.5c-f). Taken together, these evidence streams make for a compelling case for both *RNASEH2C* and *MAP3K11* as candidate genes for rs7130955. However, the mediation evidence, linear and 3D proximity and replicability of *MAP3K11* as a *cis*-eGene for rs7130955 across independent blood eQTL data sets point to this gene as the most likely candidate.

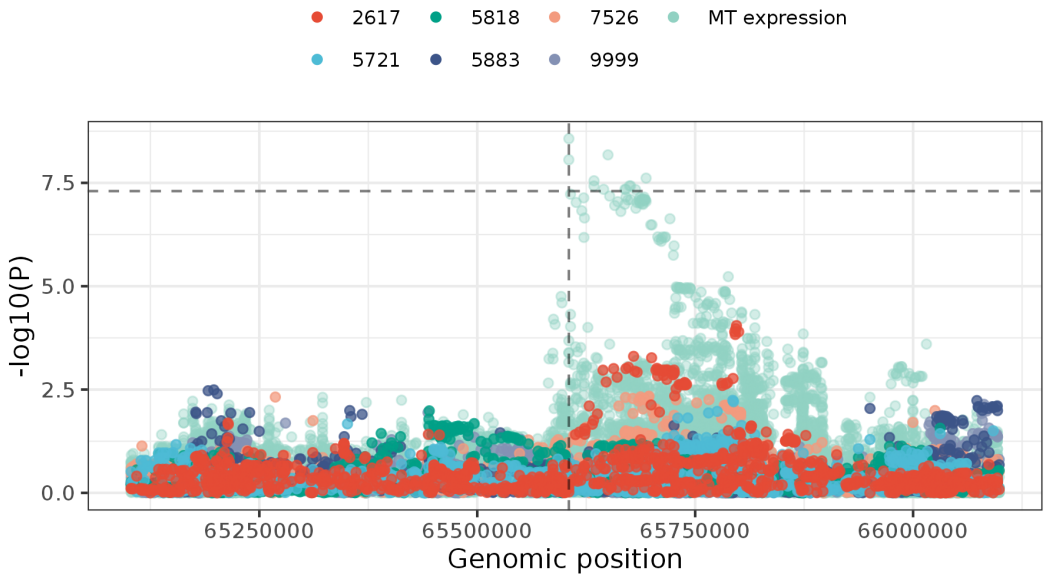

Figure 4.6: Plot to show eQTL signals (light blue) in a 1Mbp window around the variant rs7130955 (11:65605109). This has been overlaid with meQTL signals for six mitochondrial positions. Vertical line indicates the genomic position of variant rs7130955, whilst the horizontal line indicates a P-value of 5e-08.

To further understand this eQTL hit, meQTL associations for all 6 positions in a 1Mb window centred on rs7130955 were overlaid with the eQTL data (Fig.

4.6). It was interesting to note that although not reaching the G×D significance cut-off (FDR<10%), there was evidence for an association between this locus and modification rates at position 2617 (position within MT-RNR2). Thus, there is evidence arising from a second, related, transcriptional phenotype at this locus, both supporting the validity of the eQTL signal and pointing to post-transcriptional regulatory activity as a mechanism through which mtDNA gene expression is being differentially modulated between sporadic PD cases and healthy controls here.

Of the meQTLs identified, all 18 passing FDR 10% under the G×D model were associations with modification rate at mitochondrial position 5883 (Fig. 4.7a). Of these PD-specific meQTLs, two stood out as having robust beta slopes underlying their case-control differences. The first was the association between mitochondrial position 5883 and rs73044588 (12:6252835) (P=1.77e-08), characterised by a significant positive association in healthy control participants (P=4.24e-09, beta=2.86e-04) and a non-significant association in sporadic PD participants (P=0.33) (Fig. 4.7c). The V2G algorithm ranks *CD9* as the most likely eGene for rs73044588, supported by distance to the *CD9* TSS (53,120kbp) and eQTLGen data where the only eGene for rs7130955 is *CD9* (P=8.41e-7). Additionally, for 10/11 variants in linkage disequilibrium (LD, R2>0.8), *CD9* was identified as a significant eGene in the eQTLGen data. Mediation analysis between rs73044588, *CD9* and modification rate at position 5883 was non-significant (P=0.18). However, leveraging the Lohia et al., (2022) chromatin contact data, rs73044588 was found to reside in the highest contact frequency bin of the *CD9* contact map (∼1000000) (Fig. 4.7b). Here, linear and 3D proximity and replicability of *CD9* as a *cis*-eGene for rs73044588 across independent blood eQTL data sets point to *CD9* as the most likely candidate.

The second G×D interaction term meQTL of interest was the association between mitochondrial position 5883 and rs10159481 (10:37362424) (P=3.70e-07), characterised by a significant positive association in healthy control participants (P=1.99e-06, beta=1.85e-04) and a non-significant association in sporadic PD participants (P=0.075) (Fig. 4.7d). The V2G algorithm ranks *ZNF25* as the most likely

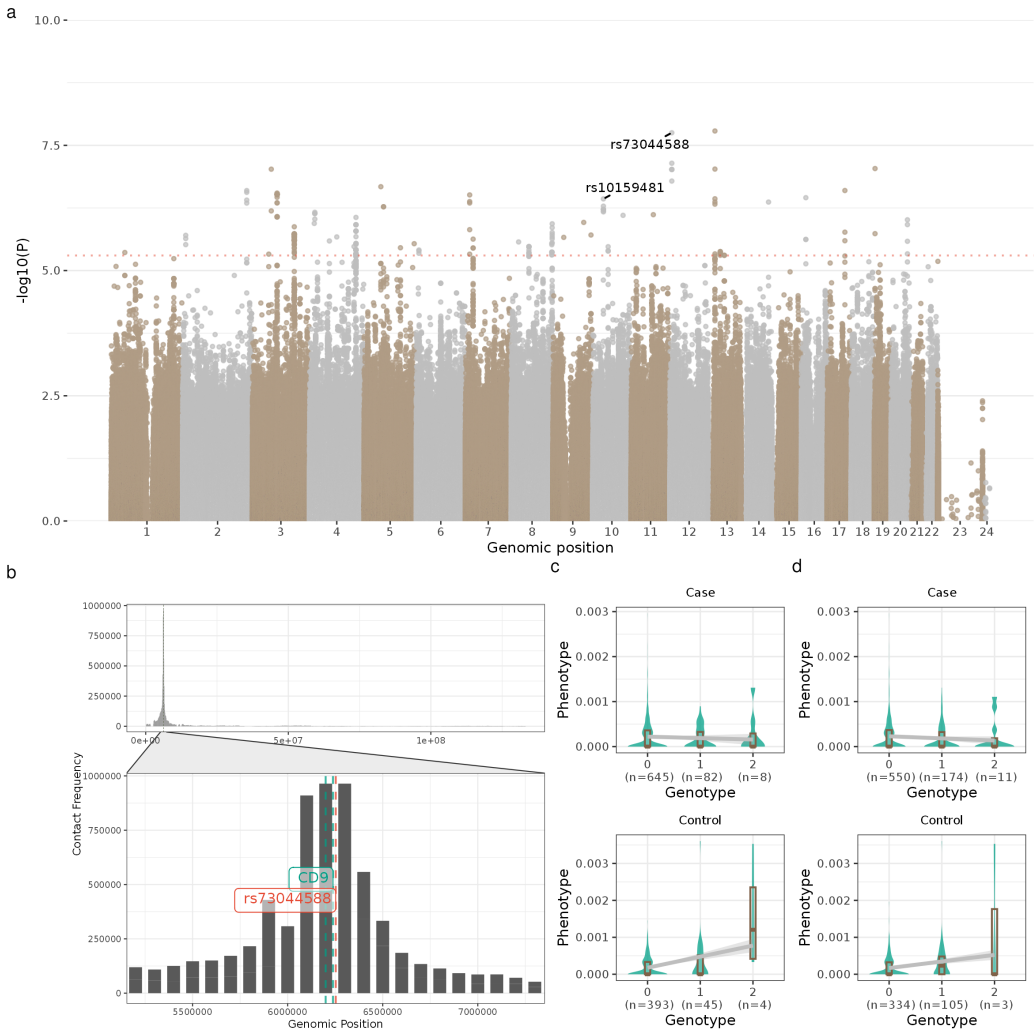

Figure 4.7: A. Manhattan plot to show all G×D associations between nuclear variants and methylation at mitochondrial position 5883. B. Plot to show Hi-C chromatin contact data between chromosome 12 and CD9 (at a resolution of 25kbp), where the gene location is indicated by two vertical green lines (start, end) and a corresponding green label. The position of variant rs73044588 is indicated by a red vertical line and corresponding red label. The zoomed area represents a 1Mbp window around the variant of interest. C. Genotype-by-phenotype distributions, where each data point represents a study participant, underlying the association between rs73044588 and methylation at mitochondrial position 5883 in sporadic PD cases (top) and healthy controls (bottom). D. Genotype-by-phenotype distributions, where each data point represents a study participant, underlying the association between rs10159481 and methylation at mitochondrial position 5883 in sporadic PD cases (top) and healthy controls (bottom).

eGene for rs10159481, supported by blood eqtl data from eQTLGen (P=2.57e-17). Additionally, for 8/8 variants in LD (R2>0.8), *ZNF25* was identified as a significant eGene in the eQTLGen data. No evidence of mediation via proximal genes was found for 5883-rs1015948, additionally, 3D proximity could not be tested as the Lohia et al., (2022) chromatin contact data did not have chromatin contact frequencies for rs10159481 eGene candidates. However, biologically, Open Targets indicates *ZNF25* as being associated wth DNA methylation and other epigenetic processes, adding to the evidence for this gene as the most likely candidate. It is important to note that a nuclear-encoded fragment of mitochondrial DNA (NUMT) was identified in proximity to variant rs10159481 (*MTRNR2L7*, 239kbp away). To ensure that this was not generating the association signal through aberrant mapping of NUMT sequences to the mitochondrial genome, during quality control, filtration for properly paired and uniquely mapped reads was carried out. However, this is only effective if the sequence similarity is low between the mtDNA-encoded gene, and so to confirm this, sequence alignment was performed between 150bp (this was the RNA-Seq library read length) windows (total=1588) of *MTRNR2L7* and the whole *MTRNR2* gene using the NCBI nucleotide BLAST tool, effectively aiming to simulate whether NUMT reads would map to the mitochondrial *MTRNR2* gene (Fig. 4.8). Considering only the full length alignments generated using this method (n=706), percentage identity (Fig. 4.8B) and number of mismatches (Fig. 4.8C) were plotted. Low sequence identity (73-90%) and high mismatch rate (13-36%) could be observed across all windows. This suggests that *MTRNR2L7* reads were unlikely to map to *MTRNR2*, meaning that it is unlikely that this meQTL signal was confounded by the nearby NUMT.

4.3.3.2 Sporadic PD case *trans*-xQTLs

To identify xQTLs that are significant in sporadic PD cases in a statistically robust manner, the PDBP data was analysed using a linear model. A significant association indicated that the effect of the genotype on the mitochondrial phenotype was significant in sporadic PD participants but without the stipulation of being significantly different in healthy controls. Two eQTLs and no meQTLs were identified

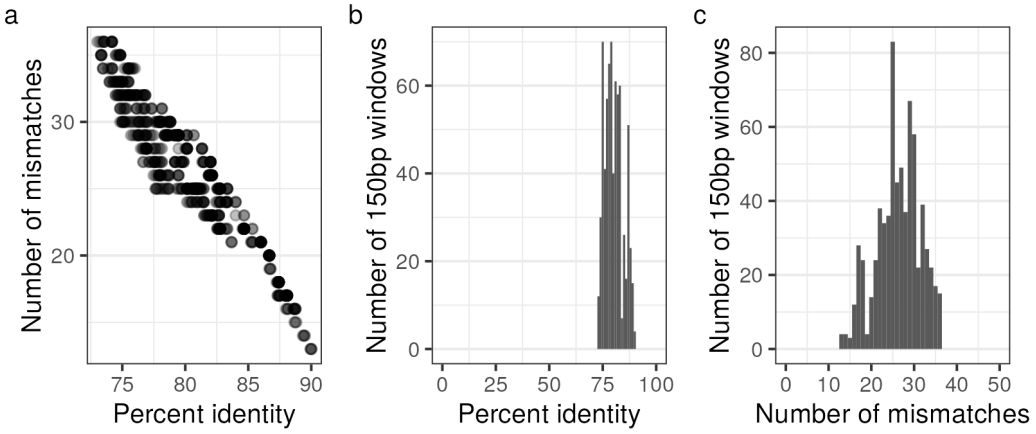

Figure 4.8: A. Percent identity against mismatch. B. Distribution of percent identity values across the 706 150bp windows. C. Distribution of mismatches across the 706 150bp windows.

in sporadic PD cases using an FDR cut-off of 5%. Of the two eQTLs identified in sporadic PD cases, one was close to significance under the G×D model using a P-value adjusted for the number of hits (P<0.05/39).

The eQTL hit identified in sporadic PD cases and trending to significance under the G×D model was the association between *MT-CO1* expression and the variant rs4907234 (2:96447202) (Fig. 4.9b). This *trans*-eQTL was characterised by a non-significant association in healthy control participants (P=0.70), a significant negative association in sporadic PD participants (P=1.82e-08) and a raw P-value under the G×D model of P=4.70e-04 (Fig. 4.9b). The V2G algorithm lists 23 candidate genes for rs4907234, of which 6 were considered as more likely candidates due to their status as a significant eGene for rs4907234 or V2G score of >0.1: *GPAT*, *ITPRIPL1*, *STARD7*, *ARID5A*, *TMEM127* and *CIAO1*. Nominally (P<0.05) significant mediation was found for *ITPRIPL1* (P=0.023) and *NCAPH* (P=0.048), with *ARID5A* near-significant (P=0.058). Leveraging the Lohia et al., (2022) chromatin contact data to provide further evidence for gene candidacy (Fig. 4.9c-h), it was observed that rs4907234 was in the highest contact frequency bin for *ARID5A* (>100000) (Fig. 4.9f). In the eQTLGen dataset, *ARID5A* is the top eGene for rs4907234 (P=1.17e-99), where *TMEM127* (P=1.38e-41) and *ITPRIPL1* (P=1.52e-11) are also listed as significant eGenes. Taken together, there is good evidence for

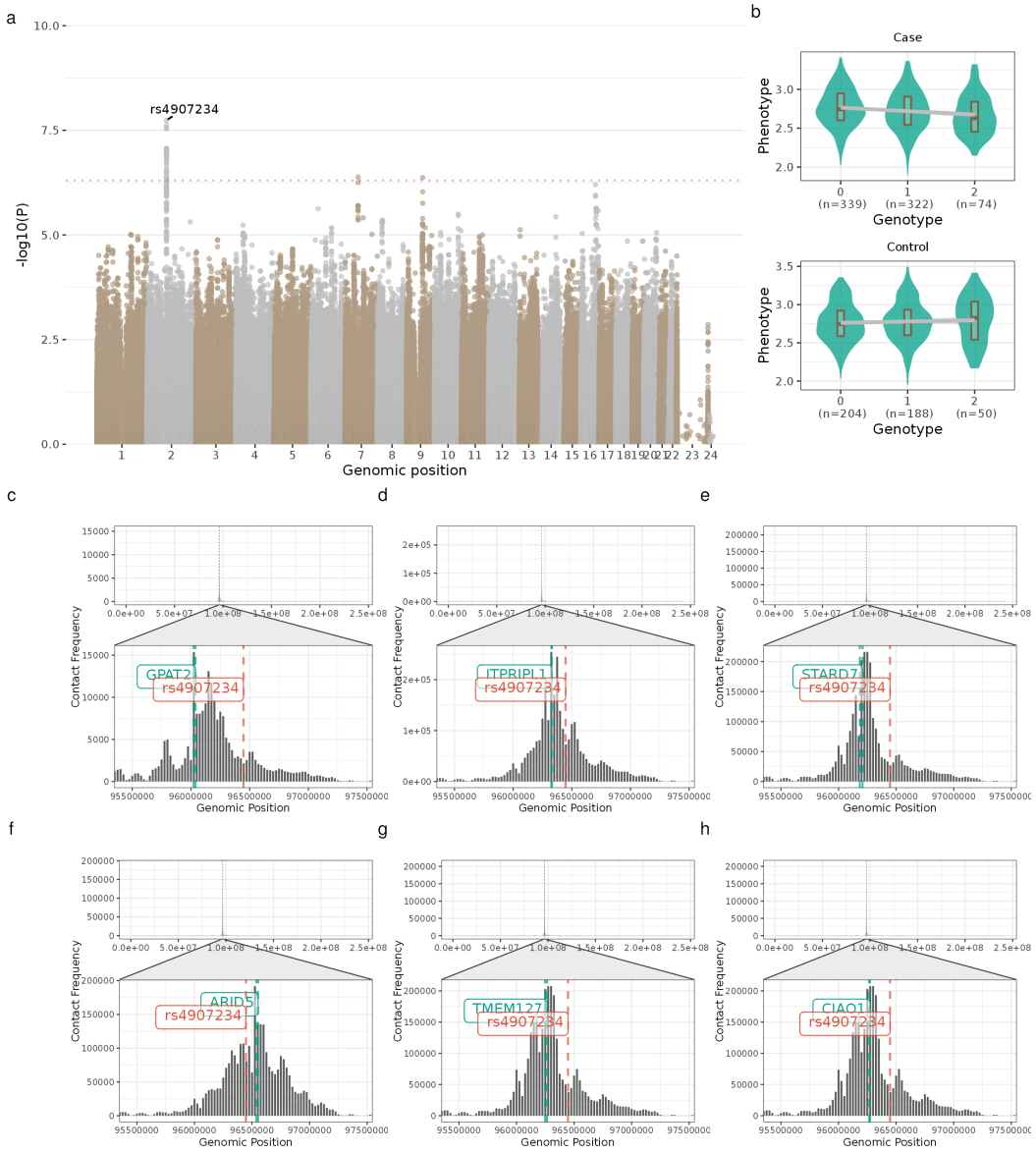

Figure 4.9: A. Manhattan plot to show all sporadic PD (non disease-specific) associations between nuclear variants and expression of the mtDNA gene *MT-CO1*. B. Genotype-by-phenotype distributions, where each data point represents a study participant, underlying the association between rs4907234 and *MT-CO1* expression in sporadic PD cases (top) and healthy controls (bottom). C-H. Plots to show Hi-C chromatin contact data between a chromosome and its host gene (at a resolution of 25kbp), where the gene location is indicated by two vertical green lines (start, end) and a corresponding green label. The position of variant rs4907234 is indicated by a red vertical line and corresponding red label. The zoomed area represents a 1Mbp window around the variant of interest.

both *ITPRIPL1* and *ARID5A* as the candidate gene for rs4907234, and as potential mediators of *MT-CO1* expression via rs4907234. *ARID5A* is a particularly interesting candidate, as its protein product is a regulator of immune responses involved in regulation of inflammatory processes. Additionally, MAPK signalling is known to regulate Arid5a through phosphorylation[203]. This could represent further support for the emerging link between nuclear regulation of mitochondrial processes and MAPK signalling in sporadic PD observed in the present study.

4.3.3.3 Healthy control *trans*-xQTLs

To identify xQTLs that are significant in healthy controls in a statistically robust manner, the PDBP data was analysed using a linear model. A significant associated indicated that the effect of the genotype on the mitochondrial phenotype was significant in healthy control participants but without the stipulation of being significantly different in sporadic PD cases. One eQTL and 10 meQTLs were identified in healthy controls using an FDR cut-off of 5%. The single control eQTL did not trend to significance under the G×D model ($P>0.05/39$), but all 10 meQTLs did trend (all $P<0.05/39$).

Of the 10 meQTLs that were identified in healthy control participants, all were trending towards significance under the G×D model, including five associations with each of position 5883 and position 7526. The single eQTL hit identified in healthy control participants was an association between *MT-CO3* and rs4724362 (7:45121704) (Fig. 4.10a). This association is characterised by a significant negative association in healthy control participants (P=1.67e-08), a negative association trending towards significance in sporadic PD participants (P=1.61e-05) and a non-significant P-value under the G×D model (P=0.38-02), meaning that this association has been detected in both healthy control and sporadic PD participants and is not significantly different, although does display differing magnitude, between sporadic PD cases and healthy controls (Fig. 4.10b). There are a number of candidate genes for rs4724362. The V2G algorithm lists 15 genes, of which 5 were considered as candidates due to their status as a significant eGene for rs4907234 or have a V2G score of >0.1: *CCM2*, *MYO1G*, *TBRG4*, *NACAD* and

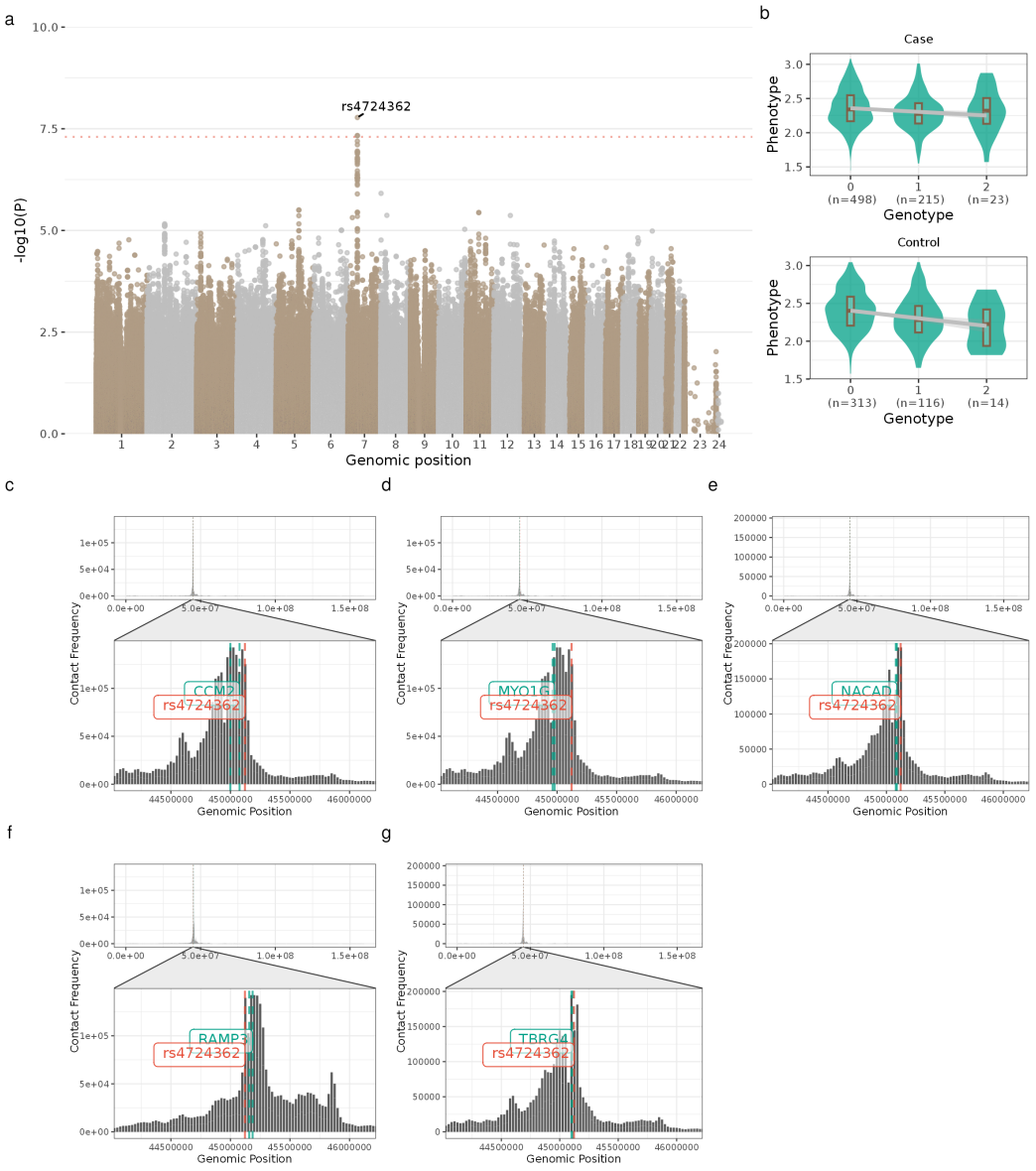

Figure 4.10: A. Manhattan plot to show all healthy control associations between nuclear variants and expression of the mtDNA gene MT-CO3. B. Genotype-by-phenotype distributions, where each data point represents a study participant, underlying the association between rs4724362 and MT-CO3 expression in sporadic PD cases (top) and healthy controls (bottom). C-G. Plots to show Hi-C chromatin contact data between a chromosome and its host gene (at a resolution of 25kbp), where the gene location is indicated by two vertical green lines (start, end) and a corresponding green label. The position of variant rs4724362 is indicated by a red vertical line and corresponding red label. The zoomed area represents a 1Mbp window around the variant of interest.

RAMP3. *TBRG4* is the nearest gene to rs4907234 has a significant negative beta in eQTLGen (P=2.14e-43), matching the direction of effect captured in the present study. *MYO1G* (P=1.24e-196) and *NACAD* (P=2.50e-46) also have significant positive beta values in eQTLGen, whereas *CCM2* has a strikingly significant negative beta (P=3.27e-310). Mediation analysis for this association finds one gene passing a Bonferroni adjusted P-value cut-off of 5%, *MYO1G* (P=1.56e-03), and four other genes passing a nominal P-value cut-off of 5%, including *CCM2* (P=1.02e-02) and interestingly, *OGDH* (P=1.83e-02) encoding the mitochondrial enzyme oxoglutarate dehydrogenase. *TBRG4* was found to be non-significant in mediation analysis (P=0.25). Unusually, the Lohia et al., (2022) chromatin contact data reveals that rs4724362 is a regions of high chromatin contact for each of the five genes, making it difficult to call a candidate using this information, although rs4724362 sits in marginally elevated contact regions for *NACAD*, *MYO1G* and *TBRG4* (Fig. 4.10c-g). The functional annotations of these genes make it most likely that *TBRG4* (*FASTKD4*) is the mediating gene here, as it is known to be involved in mitochondrial mRNA processing, regulation and stability. This result confirms that in addition to disease-specific hits, we are additionally detecting biologically relevant processes that are stable across both cases and controls. Interestingly, a SNP in LD (rs2304693, r2=0.604) with rs4724362 has been associated with mtDNA copy number, representing supporting evidence for the effect of this locus on another mitochondrial phenotype[204].

4.3.3.4 Integrating genome wide association data

Integrating GWAS and xQTL data can aid mechanistic understanding of GWAS signals, wherein the GWA study associates a genomic locus with a phenotype and the xQTLs are the association of a genomic locus with a quantitative trait (gene expression, for example). As such, colocalisation of signals can indicate that modified expression of a gene (in *cis* or *trans*) is driving the association with the phenotype at the target locus. The most recent PD risk GWAS by Nalls and colleagues was selected to test GWAS-xQTL overlap in this instance, as it is the largest and most recent PD GWAS. The aim was to understand whether any of the variants sig-

nificantly associated with mitochondrial transcriptional phenotypes in the present study were also found to be significantly associated with PD risk, thus implicating G×D association variants in PD. Of the 26 DxG associations in the present study, variants contained within 20/26 of these associations were present in the Nalls et al. GWAS summary statistics. Of these 20, two G×D associations contained variants with raw GWAS P-values below a nominal 5% cut-off. These were associations with PT modification: 5883-rs72966766 (P_{GWAS}=0.008735) and 5883-rs9869896 (P_{GWAS}=0.026820), neither of which had significant P-values when considering multiple tests. In line with this, performing colocalisation analysis did not reveal any significant colocalisations between the Nalls et al. GWAS and G×D associations detected in the present study. Given that the Nalls et al. GWAS is a meta-analysis of PD risk GWA studies, this raises the possibility that the G×D associations detected here may be more likely to modulate progression as opposed to influencing the incidence of the disease.

4.4 Discussion

This study aimed to further understand the role of mitochondrial processes in sporadic PD, and in particular, the nuclear genes and pathways that may be associated with modulating mitochondrial phenotypes specific to the disease state. To this end, this study leveraged genetic and RNA-seq data from the whole blood of 784 PD participants and 507 healthy controls, identifying genetic regulation of mitochondrial processes in sporadic PD cases, healthy controls and specific to sporadic PD for two mitochondrial transcriptional phenotypes.

Overall, eQTLs were more replicable than meQTLs, both across AMP-PD cohorts and in independent data. This was true for the single healthy control eQTL, which replicated in both datasets with high significance, and to an extent for the G×D interaction term eQTLs, for which 25% replicated (at a nominal P-value of 5%) exactly in PPMI with 33% showing partial replication with at least one replicable G×D arm. Sporadic PD case eQTLs, however, could not be replicated. Unsurprisingly, meQTLs replicated with a considerably lower rate, with 43% of healthy

control meQTLs replicating in PPMI, but none of these replicating in the independent meQTL data. G×D interaction term meQTLs also showed low to moderate replicability, with 4.8% replicating exactly in PPMI and 21% showing partial replication with at least one replicating G×D arm. There are a number of factors contributing to the moderate replication rates that we see here, particularly with respect to replication of modification rate associations. Firstly, it is important to point out that, compared to the discovery dataset, the replication dataset, PPMI has considerably less power to detect xQTLs. PPMI has 27% (n=135) of the healthy controls that PDBP has and 39% (n=302) of the sporadic PD cases that PDBP has, meaning that it likely only has the power to detect very strong signals. Secondly, although efforts were made to select a replication dataset matching the participant composition of PDBP-M0, key participant composition differences remained between the PDBP-M0 and PPMI-M36 cohorts. For example, PD case recruitment criteria, medication status, age distributions and sex ratios differ and it is possible that these factors impacted on the detection and cross-cohort replicability of *trans*-xQTLs (see Fig. 3.2 for cohort comparison). Thirdly, it is unsurprising that the meQTLs replicated less readily across cohorts and in independent data. The mitochondrial modification rates are a proxy phenotype reliant on reverse transcription error, and as such have the potential to be unstable. These proxy modification signals arise when the reverse transcriptase (RT) enzyme is blocked by base modifications and is an event that becomes less likely when higher fidelity RT enzymes are used. In the case of the AMP-PD data, the RT enzyme used was Roche Kapa HiFi polymerase (https://amp-pd.org/transcriptomics-data), which is a high fidelity RT enzyme designed to read through modified bases. In contrast, for GTEx RNA-Seq data, one of the studies included in the independent meta-analyses used for replication, the gold-standard SuperScript II RT enzyme was used, which has comparatively lower fidelity. As such, there is a high probability that the RT enzyme used in the AMP-PD library preparation ultimately limited the use of mismatch proportion as a proxy to detect mitochondrial genomic modifications, resulting in the low meQTL replicability observed here.

In addition to using replication to assess the robustness of the associations detected, efforts were made to demonstrate that biologically relevant associations were being detected in sporadic PD cases that were also detected in healthy control participants. It was shown that the most stable high-significance xQTLs were associated with nDNA-encoded mitochondrial genes and processes. This indicates that associations being detected in sporadic PD participants were related to normal, healthy mitochondrial processes, suggesting that these associations were not simply a consequence of mitochondrial dysfunction. An example of this, detected at FDR$<$0.05 in healthy control participants was the association between rs4724362 and *MT-CO3* expression which is also detected with high but sub-threshhold significance in cases (P=1.61e-05) and whose likely candidate gene is *TBRG4* (also known as *FASTKD4*). *TBRG4*, transforming growth factor beta regulator 4, is known to be involved in mitochondrial mRNA processing, regulation of mitochondrial mRNA stability and has been found to localise to the mitochondrial matrix. Thus, this is one example of an association which is detected consistently across diagnoses and is linked to mitochondrial biology. Notably, this association was also strongly replicated in the independent healthy control data, supporting its validity.

The top G$\times$D interaction term eQTL was an association between *MT-ND6* expression and the nuclear variant rs7130955. This association was characterised by a significantly different association in sporadic PD cases compared with controls, where the association was positive in healthy control participants with a marginally weaker negative association in sporadic PD participants (Fig. 4.5). Multiple streams of evidence (proximity, mediation, independent whole blood *cis*-eQTL data, HiC data) pointed to *MAP3K11* as the *cis*-mediator of this association. As such, the model proposed here is that rs7130955 has an opposing effect in healthy control participants compared to sporadic PD case participants, for which a potential biological mechanism may be that the nuclear genetic variant contributes to disease susceptibility or progression through a *cis* effect on the *MAP3K11* gene. *MAP3K11*, mitogen-activated protein kinase kinase kinase 11, encodes MLK3, a member of the serine/threonine kinase family known to preferentially activate MAPK8/JNK

kinase and positively regulate the JNK signalling pathway which is responsible for promotion of apoptosis[205,206]. Interestingly, *MAP3K11* has previously been explored as a target for slowing PD progression. The MLK3 inhibitor CEP-1347 was found to prevent motor deficits and neuronal degradation in murine MPTP (neurotoxic to dopaminergic neurons) models and following promising pre-clinical results, reached phase II trials for PD in 2006 (n=806)[207,208,209]. Explanations for trial failure included the inability of MLK inhibitors alone to slow PD progression, off-target effects including the inhibition of other kinases, unsuccessful accumulation of CEP-1347 to therapeutic levels due to its biochemical characteristics and insufficient blood-brain barrier penetrance[207]. Building on this work, the Kline et al., (2019) study utilised murine models with MTPT-induced nigrostriatal dopaminergic pathology, showing that a novel 'second generation' MLK3 inhibitor with improved blood-brain barrier penetrance and higher specificity for MLK3, CLFB-1134, inhibited phosphorylation of MLK3 targets and protected against neurotoxin-induced nigral dopaminergic neuron loss[205]. The same study concluded that inhibition of MLK3 may be a valid target for future work investigating PD treatments[205]. Overlapping meQTL (association with PT modification at position 2617) and eQTL signals (Fig. 4.6) builds on confidence that association signals observed here are genuine and robust, but untangling causal processes here is challenging given the observation of opposite effects in cases and controls. As such, the results of the present study may point to a mitochondrial role for *MAP3K11*, in particular suggesting its potential relevance to sporadic PD aetiology.

Two notable G×D interaction term meQTLs identified were associations with modification rate at position 5883, a tRNA p9 site between the protein coding genes *MT-ND2* and *MT-CO1*. The first is an association with variant rs73044588, for which the candidate gene most highly supported is *CD9* (*TSPAN29*), encoding a member of the tetraspanin family. This is characterised by a positive association in controls, where the variant is associated with higher modification rates. In sporadic PD cases, there is no significant association meaning that the variant does not appear to confer this effect. Higher modification at 5883, a p9 site, is thought to sta-

bilise the tRNA transcript structure, allowing for RNA cleavage and the release of gene products, ultimately increasing the availability of OXPHOS components[189]. As such, the model presented by the G×D interaction term meQTL here is that in controls, the variant - putatively mediated by *CD9* - is associated with higher modification of 5883, resulting in greater availability of OXPHOS components. Although no evidence exists for a direct role for *CD9* in PT modification processes, a number of studies have uncovered a role for *CD9* in mitochondrial-lysosomal pathways. In melanoma cell lines, for example, *CD9* was found to regulate mitochondrial clearance and it is thought that this operates through a role in late endosome maturation and incorporating cargo into early endosomes[210,211]. Another study showed that *CD9*/*CD81* double knockout mice exhibited accelerated ageing, including a shortened lifespan, linked to reduced *SIRT1* expression which is a core regulator of cellular senescence and inflammation pathways[212]. Interestingly, *SIRT1* is known to regulate autophagy, mitochondrial function and inhibit oxidative stress[213]. A number of studies have also demonstrated a role for *SIRT1* in PD, through regulation of neuroinflammation[213]. Thus, this study provides evidence for *CD9* mediating, or perhaps indirectly mediating via interaction with *SIRT1*, for example, an effect on mitochondrial transcription through PT methylation of mitochondrial RNA. It is also possible that modification rate responds dynamically to processes driven by *CD9*. Further mechanistic work is necessary to understand the precise mechanisms driving this association.

It is important to acknowledge the limitations of this study and the resulting models. Firstly, the tissue utilised in this study was whole blood, which is not the tissue in which the disease primarily exerts its symptomatic effects. However, growing evidence points to the blood as a useful tissue for the identification of PD biomarkers, with known PD pathophysiology being detected in blood, for example alpha synuclein species, lysosomal enzymes and amyloid and tau pathology[214]. Additionally, recent work has demonstrated clear transcriptomic and methylomic shifts in whole blood samples from individuals with PD[215,216]. Despite this, important cell type specific associations may not be detected using blood alone. To

remedy this, it is hoped that large sporadic PD cohorts will be generated in future. Post-mortem brain tissue would have utility in this regard, although this too is associated with a number of issues (post-mortem interval, RNA degradation, loss of focal cell types). Other large datasets based on data collected from iPSC experiments, for example, may also allow exploration of mitochondrial-nuclear associations in specific cell types. Secondly, despite usage of a large whole-blood dataset to gain power for the detection of *trans*-xQTLs, the effect sizes of these are inherently small, making them difficult to detect. Future expansion of whole blood sporadic PD/ healthy control cohorts may permit the detection of additional small effect mitochondrial-nuclear *trans*-xQTLs and will likely reveal sub-threshhold hits as significant. Thirdly, despite rigorous attention to variant-to-gene assignment, using a number of strategies beyond the standardly applied 'nearest gene', it remains a possibility that candidate nuclear genes have been mis-identified. Furthermore, for the associations identified, only a single *cis*-mediator of each association was considered. It has been shown that *trans*-eQTLs are more likely than randomly selected SNPs to associate with more than one *cis*-gene[217]. An interesting direction for future work could be to move beyond the single *cis*-mediator assumption and to instead look at whether nuclear variants associated with mitochondrial phenotypes do this through multiple *cis*-mediator effects. This may help to identify additional nuclear genes and pathways which act to modulate mtDNA gene expression and PT modification. Finally, the present study is limited to the detection of statistically significant *trans*-xQTLs, and while this is a crucial step for the identification of important biological contributors to sporadic PD, to further understand and unpick the models proposed here, detailed functional work is required.

In conclusion, using QTL mapping this study finds that nuclear genetic regulation of mtDNA-encoded genes is perturbed in PD, implicating *MAP3K11* and *CD9* in disease pathways that also operate through modulation of mitochondrial processes. These represent novel molecular features with the potential to reveal new biology and to enable development of therapeutic targets for sporadic PD.

Chapter 5

Conclusions and future directions

The prevalence of adult NDs is rising, but we still lack a full understanding of the pathogenesis of these conditions. We also lack biomarkers that would allow early and reliable diagnosis and there are currently limited interventions that prevent or impede progression. To solve these problems, efforts will be multi-disciplinary, will leverage rapidly advancing 'omics technologies including single-cell and spatial RNA-Seq, stem cell technologies — such as induced pluripotent stem cells (iPSCs) and organoids — and imaging techniques.

Mitochondrial dysfunction has been implicated in the pathogenesis of AD and PD, the two most common NDs and as such, this thesis aimed to contribute to the wider understanding of these diseases through studying one potential route to mitochondrial dysfunction. Ample evidence points to coordination between the mitochondrial and nuclear genomes being essential to normal mitochondrial function, but formal analysis of this in specific NDs is absent from the literature. As such, this thesis aimed to study this and did so through the lens of transcriptomics.

Each results chapter of the thesis (two, three and four) provides a discussion of its findings, however, viewed as a whole, three main themes emerge and require further discussion: (i) mitochondrial-nuclear coordination and cell type specificity, (ii) altered mitochondrial-nuclear coordination and NDs and, (iii) the timing of mitochondrial dysfunction in ND aetiology.

5.0.1 Mitochondrial-nuclear coordination and cell type specificity

The first theme emerging from this body of work is that of cell type specificity. Work in chapter two found that mitochondrial-nuclear correlation profiles were regionally distinct in healthy CNS tissue. Hypothesising that this was driven by regional cell type composition, cell type specificity values derived from mouse brain scRNA-seq data were used to estimate the enrichment of cell type markers in mitochondrial-nuclear pairs with extreme expression correlation (high negative and high positive). Extreme positive and extreme negative pairs being enriched for different cell types pointed towards cell type composition as the likely driver of cross-CNS differences. Further to this, it was observed that mitochondrial-nuclear correlations with high cross-CNS variance were enriched for synaptic genes, pointing to synaptic pathways as harbouring variable mitochondrial-nuclear coordination and indicating potential cell type or sub-celltype specificity. At the time of analysis (2019/20), these results could not be followed up using single-cell RNA-seq (scRNA-seq) data because datasets of the appropriate scale, design or type (human, multiple CNS tissues, AD/PD case-control) were not publicly available.

Since this time, single-cell technology has rapidly advanced. Increasingly large and comprehensive datasets have been generated, holding promise for the interrogation of cell type-specific processes. However, the prevalence of single nuclear RNA-Seq (snRNA-Seq) over single cell RNA-Seq (scRNA-Seq) datasets is a distinct limitation in terms of the present work as mtDNA-encoded RNA is not sequenced in snRNA-Seq. As such, to study mitochondrial-nuclear coordination in human cell types, scRNA-Seq data derived from human iPSCs is one potential avenue. Another is bulk RNA-Seq from purified cell cultures. Studying mitochondrial-nuclear coordination in such a dataset would enable a fuller understanding of its cell type specificity, allowing elucidation of the cell types where such coordination is most important and ultimately identifying those that may be especially vulnerable to its dysfunction.

Selective neuronal vulnerability is a present focus of ND research, and much

is still to be understood. Given that clinical manifestation of NDs are often a direct reflection of the specific cell populations that degenerate, this could be of key interest[218]. An understanding of this would yield insights into the molecular mechanisms underlying NDs. In chapter 2, genes with variable mitochondrial-nuclear correlation were enriched for nuclear genes involved in synaptic function. This may reflect neuronal subtype-specific variation arising from energetic specialisation, perhaps indicating a fine-tuning of mitochondrial-nuclear coordination. Indeed, accumulating evidence points to cell type specific mitochondrial specialisation wherein morphological differences between the mitochondria of neuronal subtypes has been identified[219,126]. While the field of mitochondrial specialisation is still relatively new, as early as 2005 it was noticed that sub-populations of isolated mitochondria differentially adapted in response to a metabolic challenge[220]. Building on this, it is likely that the brain and its composite cell types would be a particularly interesting place to look at mitochondrial specialisation for a number of reasons: (i) neuronal subtypes are known to be highly heterogeneous both functionally and morphologically[128,221], (ii) the nuclei are often highly distal from synapses (soma-synapse distances range from a millimetre up to a meter)[222], and (iii) many brain cell types are highly energetically demanding[223]. These features suggest that mitochondrial-nuclear coordination might be both specifically important and specifically challenging in neuronal cells. As such, it is possible that adaptations to these challenges may manifest as intracellular or intercellular mitochondrial specialisation. Technologies that can sample mitochondria from subcellular regions are required to study this, for example, nanoscale tweezers which have recently been used to successfully extract a single mitochondrion[224] may elucidate heterogeneity of protein localisation, morphology which could contribute to further understanding long-distance mitochondrial-nuclear communication that occurs in CNS cell types. Ultimately, this granularity of understanding may provide insights into the selective vulnerability of cell types and in turn ND aetiology.

5.0.2 Altered mitochondrial-nuclear coordination and neurodegenerative disease

Despite the multiple paths to neurodegeneration present both within and across disorders, NDs share certain cellular hallmarks, including DNA damage, oxidative damage, impaired autophagy, dysregulated bioenergetics and mitochondrial dysfunction, which culminate in neuronal cell death and neurodegeneration[225]. This manifests as the sensory, cognitive and movement deficits that are characteristic of NDs[225]. In order to develop and target interventions, an understanding of the processes underpinning the dysfunction of these hallmarks is required. This thesis focuses on one of these hallmarks - mitochondrial dysfunction - for which the causality and progression of the dysfunction have not been fully characterised. To achieve this, the complex cellular architecture in which mitochondria operate must be understood and looking to its relationship with the nucleus — its essential functional partner — is fertile ground for this.

In chapter two, mitochondrial-nuclear coordination is assayed by examining the correlation between mitochondrial and nuclear gene expression, characterising coordination across the healthy CNS and in the cortex of individuals diagnosed with AD. In the AD brain, significant mitochondrial-nuclear correlation differences were observed compared with healthy controls. Within this, gene pairs with a highly disparate correlation in cases compared to controls were enriched for nuclear genes in synaptic and post-synaptic pathways. In a set of genes relating to adult onset NDs, it was found that there was a stronger correlation with the mitochondrial genome than could be expected by chance, implicating NDs more generally. The correlation assay was supported by QTL mapping work (in chapter 4) which revealed the presence of nuclear variants differentially modulating mitochondrial phenotypes in individuals with sporadic PD compared to healthy control individuals. Chapter two then assayed mitochondrial expression and PT modification profiles on different nuclear mutational backgrounds. Distinct mitochondrial profiles were observed in monogenic PD cases versus healthy controls, pointing to a role for the nuclear genome in mitochondrial transcriptional changes in the disease state. Taken together, the

findings of this thesis support a role for mitochondrial-nuclear coordination in PD and AD but also point to a role for this relationship in NDs more generally.

Mitochondrial-nuclear coordination is often referred to as 'cross-talk', a term attempting to capture a complex scenario where two organelles coordinate to dynamically respond to and manage, for example, energy demands. Cross-talk is something that operates at many organisational levels (genetic, transcription, translation, protein, chemical) and involves bi-directional interaction (anterograde and retrograde), feedback loops, other organelles and responses to physiological and environmental cues[225]. One example of a process that could be better understood through characterisation of mitochondrial-nuclear cross-talk involvement is the mito-lysosomal axis. The lysosomal pathway is implicated in AD in this work by lysosomal gene enrichment in mitochondrial-nuclear gene pairs displaying disparate correlation magnitudes in individuals with AD compared to controls (chapter two). Lysosomal dysfunction has been linked to several other NDs, including PD, frontotemporal dementia (FTD) and amyotrophic lateral sclerosis (ALS), suggesting the existence of a common dysfunction between these diseases[226]. Lysosomes play an essential role in mitophagy, which is a core component of mitochondrial quality control, where they act to degrade damaged or dysfunctional mitochondria through binding to encapsulating autophagocytes and releasing acid hydrolase enzymes[227]. Interestingly, more complex roles for lysosomes than their common portrayal as cellular 'rubbish bags' have been observed recently, which may present further avenues for understanding of their dysfunction, however, these will not be considered here.

It has been established that lysosomal biogenesis can be induced by acute mitochondrial respiratory chain stress, and that this involves the TFEB transcription factor family, AMPK signalling and calcineurin signalling, during which, relocation of TFEB to the nucleus occurs[228]. However, the same study identified that in murine models of chronic mitochondrial dysfunction, the transcriptional program of lysosomal biogenesis was repressed, noting the presence of an opposite response in acute versus chronic mitochondrial stress and finding that this was consistent

across multiple *in vitro* models[228]. This work illuminates some intricacies of the mito-lysosomal pathway, but many aspects remain cryptic. For example, in the chronic stress response, the mechanism by which the transcriptional triggers for lysosomal biogenesis are deactivated is not clear. Additionally, in the acute stress response, exactly how AMPK is activated is not understood. To elucidate these mechanisms, understanding mitochondrial-nuclear communication may be crucial. The nucleus is a clear mediator of the mito-lysosomal interaction, encoding all lysosomal and the majority of mitochondrial components. Additionally, the nucleus provides all components of the relevant signalling cascades, acts as the site of TFEB localisation and is the site of lysosomal biogenesis transcriptional repression. It would seem that cross-talk between the nucleus and mitochondria is an essential contributor to mito-lysosomal interaction and may go some way explaining the missing links. Under this model, potential avenues for further interrogation of mitochondrial-nuclear cross-talk may include further exploration of: (i) reterograde (mitochondria to nucleus) signalling, which is known to involve chemical signals of mitochondrial stress such as ROS, calcium and ATP/ADP ratios, (ii) anterograde dynamic provision of nDNA-encoded proteins required mitochondrial function (e.g. OXPHOS components, transcriptional/translational machinery, or (iii) nuclear initiation of signalling pathways that affect mitochondria (for example, initiation of mitophagy). In the case of the mito-lysosomal pathway, it is possible that any three of these mitochondrial-nuclear mechanisms are involved. As such, further understanding of this relationship may shed light on the close interactions between mitochondria and lysosomes which has utility for understanding the pathogenesis of NDs in which lysosomal dysfunction has been implicated.

The lysosomal pathway is one example where understanding mitochondrial-nuclear coordination may provide insights, but more generally, a deeper understanding of the molecular mechanisms that converge across NDs such as this could help to explain the observed clinical and pathological similarities between them. To this end, future directions will involve interdisciplinary work leveraging multiple established and emerging technologies, including therapeutic target identification using

'omics technologies paired with granular functional studies using iPSCs, organelle isolation and CRISPRi/CRISPRa screens[226]. Particularly in NDs, for which selective vulnerability is a feature, single cell 'omics may be pivotal for identifying cell type-specific signatures.

5.0.3 The timing of mitochondrial dysfunction in Parkinson's disease aetiology

Until now, we have been attempting to further understand mitochondrial dysfunction through the lens of mitochondrial-nuclear coordination, and asking whether its failure could contribute to mitochondrial dysfunction in NDs. Work in this thesis suggests that failure of this relationship is a feature of NDs, and for PD specifically, many studies have implicated mitochondrial failure in disease pathogenesis. However, the initial trigger for neurodegeneration has not been identified[229]. As such, it remains unclear whether mitochondrial dysfunction is a cause or consequence of disease, and whether it arises early or late in the disease trajectory. An understanding of the order of events in NDs is crucial for the development of interventions seeking to halt or slow progression and as such, it is of value to know when mitochondrial dysfunction arises. Work in chapter three highlights that in familial PD, profiles of mitochondrial OXPHOS expression were more similar between asymptomatic and symptomatic participants than healthy control participants. This finding may point to mitochondrial dysfunction becoming established prior to symptom onset and consequently that arises early in the disease trajectory. Indeed recent work in mice found that damaging complex I alone is sufficient to trigger axonal, motor learning, fine motor and dopamine release deficits characteristic of PD, resulting in a levodopa responsive disease[230]. The authors state that this goes against the current paradigm, signalling a shift in thinking towards mitochondria as a driver of PD. Furthermore, recent work tracked the aggregation of α-synuclein in neurons and found that oligomerisation occurred at mitochondrial membranes and that this led to neuronal toxicity[231]. However, disentangling the causes and the effects of neurodegeneration are difficult, particularly with regard to mitochondria where many mitochondrial pathways implicated in NDs have diverse cellular roles. There

are many facets to mitochondrial dysfunction and levels at which it may act, potentially with temporal differences: mitochondrial respiratory chain, mtDNA mutation, mitochondrial dynamics, distribution and transport, mitophagy, biogenesis, protein aggregation, calcium regulation and ROS production have all been highlighted[229]. Understanding more about each of these processes leveraging diverse techniques will help us to understand how and exactly when mitochondrial dysfunction occurs, paving the way for the targeting of interventions to the right individuals at the right time.

5.0.4 Concluding remarks

All considered, the results presented in this thesis support evidence for the prominent role of mitochondria in ND, implicating shifts or failure of the mitochondrial-nuclear relationship in ND pathogenesis. The mitochondrial-nuclear relationship was found to be CNS tissue specific, cell type specific and to be modulated in AD and PD. In sporadic PD, for which the aetiology remains elusive, this work identifies that one feature of mitochondrial-nuclear dysfunction may be the presence of nuclear variants differentially modulating mitochondrial transcription in sporadic PD compared to healthy controls. The implications of this work lie in both furthering our understanding of how exactly mitochondrial dysfunction occurs and further to this, how patients can be identified, stratified and treated for PD and other NDs for which mitochondrial dysfunction is a hallmark.

Appendix A

Supplementary tables

gene_set	p_value	term_size	query_size	intersection_size	precision	recall	term_id	source	term_name	effective_domain_size	source_order
high_var_GE_pos	0.011836	40	128	5	0.039063	0.125	WP:WP4222	WP	Phosphodiesterases in neuronal function	14686	475
high_var_GE_neg	7.37E-06	528	113	25	0.221239	0.047348	GO:0098916	GO:BP	anterograde trans-synaptic signaling	14686	22515
high_var_GE_neg	7.37E-06	528	113	25	0.221239	0.047348	GO:0007268	GO:BP	chemical synaptic transmission	14686	3194
high_var_GE_neg	9.59E-06	537	113	25	0.221239	0.046555	GO:0099537	GO:BP	trans-synaptic signaling	14686	22685
high_var_GE_neg	1.65E-05	551	113	25	0.221239	0.045372	GO:0099536	GO:BP	synaptic signaling	14686	22684
high_var_GE_neg	0.018093	4	67	3	0.044776	0.75	GO:0098971	GO:BP	anterograde dendritic transport of neurotransmitter receptor complex	14686	22549
high_var_GE_neg	0.030041	368	80	13	0.1625	0.035326	GO:0050808	GO:BP	synapse organization	14686	15584
high_var_GE_neg	1.42E-06	327	139	21	0.151079	0.06422	GO:0098978	GO:CC	glutamatergic synapse	14686	3732
high_var_GE_neg	2.66E-06	1109	118	36	0.305085	0.032462	GO:0045202	GO:CC	synapse	14686	2468
high_var_GE_neg	0.0007	1706	96	36	0.375	0.021102	GO:0030054	GO:CC	cell junction	14686	1045
high_var_GE_neg	0.003685	332	139	17	0.122302	0.051205	GO:0099572	GO:CC	postsynaptic specialization	14686	3825
high_var_GE_neg	0.007746	316	139	16	0.115108	0.050633	GO:0014069	GO:CC	postsynaptic density	14686	842
high_var_GE_neg	0.009804	321	139	16	0.115108	0.049844	GO:0032279	GO:CC	asymmetric synapse	14686	1529
high_var_GE_neg	0.017186	103	137	10	0.072993	0.097087	GO:0099634	GO:CC	postsynaptic specialization membrane	14686	3831
high_var_GE_neg	0.020715	546	92	17	0.184783	0.031136	GO:0098794	GO:CC	postsynapse	14686	3659
high_var_GE_neg	0.020754	85	137	9	0.065693	0.105882	GO:0098839	GO:CC	postsynaptic density membrane	14686	3681
high_var_GE_neg	0.023141	339	139	16	0.115108	0.047198	GO:0098984	GO:CC	neuron to neuron synapse	14686	3738
high_var_GE_neg	0.046917	97	118	9	0.076271	0.092784	GO:0098936	GO:CC	intrinsic component of postsynaptic membrane	14686	3715
high_var_GE_neg	0.00192	82	129	7	0.054264	0.085366	KEGG:04713	KEGG	Circadian entrainment	14686	324
high_var_GE_neg	0.019004	174	131	9	0.068702	0.051724	KEGG:04014	KEGG	Ras signaling pathway	14686	225
high_var_GE_neg	0.020951	161	72	6	0.083333	0.037267	KEGG:04360	KEGG	Axon guidance	14686	282
high_var_GE_neg	0.034015	115	107	6	0.056075	0.052174	KEGG:04728	KEGG	Dopaminergic synapse	14686	334
high_var_GE_neg	0.000308	294	135	16	0.118519	0.054422	REAC:R-HSA-112316	REAC	Neuronal System	14686	1420
high_var_GE_neg	0.030563	84	72	5	0.069444	0.059524	REAC:R-HSA-2682334	REAC	EPH-Ephrin signaling	14686	735
high_var_GE_neg	0.004519	149	131	9	0.068702	0.060403	WP:WP4223	WP	Ras Signaling	14686	476
high_var_GE_neg	0.040334	30	61	3	0.04918	0.1	WP:WP4875	WP	Disruption of postsynaptic signalling by CNV	14686	591
low_var_GE_pos	0.000812	8	49	3	0.061224	0.375	REAC:R-HSA-195399	REAC	VEGF binds to VEGFR leading to receptor dimerization	14686	2324
low_var_GE_pos	0.000812	8	49	3	0.061224	0.375	REAC:R-HSA-194313	REAC	VEGF ligand-receptor interactions	14686	2325
low_var_GE_pos	0.023974	67	303	9	0.029703	0.134328	REAC:R-HSA-216083	REAC	Integrin cell surface interactions	14686	1090
low_var_GE_pos	0.024101	4	46	2	0.043478	0.5	REAC:R-HSA-194306	REAC	Neurophilin interactions with VEGF and VEGFR	14686	1421
low_var_GE_pos	0.018125	15	88	3	0.034091	0.2	WP:WP3967	WP	miR-509-3p alteration of YAP1/ECM axis	14686	104
low_var_GE_pos	0.02383	18	239	4	0.016736	0.222222	WP:WP4533	WP	Transcription co-factors SKI and SKIL protein partners	14686	540
low_var_GE_neg	0.007723	850	357	50	0.140056	0.058824	GO:0006396	GO:BP	RNA processing	14686	2421
low_var_GE_neg	0.03344	629	357	33	0.092437	0.052464	REAC:R-HSA-8953854	REAC	Metabolism of RNA	14686	1265

Table A.1: Table detailing all gProfiler enrichments for four cross-CNS variance-defined gene lists. Gene lists were defined as follows: high variance positive (high_var_GE_pos), high variance negative (high_var_GE_neg), low variance positive (low_var_GE_pos) and low variance negative (low_var_GE_neg).

Participant category	Inclusion criteria	Exclusion criteria
PD	1. Male or female age 30 years or older at time of PD diagnosis	1. Currently taking levodopa, dopamine agonists, MAO-B inhibitors, amantadine or other PD medication

Participant category	Inclusion criteria	Exclusion criteria
	2. At least two of: resting tremor, bradykinesia, rigidity (must have either resting tremor or bradykinesia); OR either asymmetric resting tremor or asymmetric bradykinesia	2. Has taken levodopa, dopamine agonists, MAO-B inhibitors or amantadine within 60 days of Baseline
	3. A diagnosis of Parkinson disease for 2 years or less at Screening	3. Has taken levodopa or dopamine agonists prior to Baseline for more than a total of 60 days
	4. Hoehn and Yahr stage I or II at Baseline	4. Received drugs that might interfere with dopamine transporter SPECT imaging (i.e. neuroleptics) within 6 months of Screening
	5. A SPECT scan result consistent with dopamine transporter deficit	5. Current treatment with anticoagulants that might preclude safe completion of the lumbar puncture
	6. Not expected to require PD medication within at least 6 months from Baseline.	6. Condition that precludes the safe performance of routine lumbar puncture
		7. Use of investigational drugs or devices within 60 days prior to Baseline
Healthy control	1. Male or female age 30 years or older at Screening	1. Current or active clinically significant neurological disorder
		2. First degree relative with idiopathic PD (parent, sibling, child)

Participant category	Inclusion criteria	Exclusion criteria
		3. Montreal Cognitive Assessment (MoCA) test for dementia score <26 (i.e. non-normal) 4. Received drugs that might interfere with dopamine transporter SPECT imaging (i.e. neuroleptics) within 6 months of Screening 5. Current treatment with anticoagulants that might preclude safe completion of the lumbar puncture 6. Condition that precludes the safe performance of routine lumbar puncture 7. Use of investigational drugs or devices within 60 days prior to Baseline
SWEDD	1. All inclusion criteria for 'PD' patients apply, except SPECT scan must show no evidence of dopamine transporter defecit	1. All exclusion criteria for 'PD' patients apply
Prodromal	1. All inclusion criteria for 'PD' patients apply, expect these patients are aged 60 or older and have a REM sleep disorder	1. All exclusion criteria for 'PD' patients apply
Genetic cohort (affected)	1. Male or female age 18 years or older	1. Current treatment with anticoagulants (e.g. coumadin, heparin) that might preclude safe completion of the lumbar puncture

Participant category	Inclusion criteria	Exclusion criteria
	2. At least two of: resting tremor, bradykinesia, rigidity (must have either resting tremor or bradykinesia); OR either asymmetric resting tremor or asymmetric bradykinesia	2. Condition that precludes the safe performance of routine lumbar puncture, such as prohibitive lumbar spinal disease, bleeding diathesis, or clinically significant coagulopathy or thrombocytopenia
	3. A diagnosis of Parkinson disease for 7 years or less at Screening	
	4. Hoehn and Yahr stage <IV at Baseline	
	5. Willingness to undergo genetic testing and to be informed of genetic testing results	
	6. Confirmation of mutation in LRRK2, GBA or SNCA	
	7. For subjects taking any drugs that might interfere with dopamine transporter SPECT imaging (Neuroleptics, metoclopramide, alpha methyldopa, methylphenidate, reserpine, or amphetamine derivative) must be willing and able from a medical standpoint to hold the medication for at least 5 half-lives prior to screening DatSCANTM imaging	

Participant category	Inclusion criteria	Exclusion criteria
Genetic cohort (unaffected)	1. Male or female age 45 years or older at baseline with a LRRK2 or GBA mutation and/or a first degree relative with a LRRK2 or GBA mutation OR Male or female age 30 years or older at baseline with a SNCA mutation and/or a first degree relative with a SNCA mutation	1. A clinical diagnosis of PD
	1. Unaffected subjects at high risk of LRRK2, GBA or SNCA mutation due to first degree relative with a LRRK2, GBA or SNCA mutation may choose either to be informed of the results or remain unaware of the results	2. Current treatment with anticoagulants (e.g. coumadin, heparin) that might preclude safe completion of the lumbar puncture
	2. Unaffected subjects from an ethnic or geographic group known to have relatively high risk of LRRK2, GBA or SNCA mutation (such as people of Ashkenazi Jewish or Basques descent) and who have a family member (either alive or deceased) who has/had PD must be willing to be informed of their own testing results	3. Condition that precludes the safe performance of routine lumbar puncture, such as prohibitive lumbar spinal disease, bleeding diathesis, or clinically significant coagulopathy or thrombocytopenia
	3. Willingness to undergo genetic testing	

Participant category	Inclusion criteria	Exclusion criteria
	4. For subjects taking any of the following drugs that might interfere with dopamine transporter SPECT imaging (Neuroleptics, metoclopramide, alpha methyldopa, methylphenidate, reserpine, or amphetamine derivative) must be willing and able from a medical standpoint to hold the medication for at least 5 half-lives prior to DatSCAN imaging	
Genetic registry	1. Male or female age 18 years or older 2. Individual with a LRRK2, GBA or SNCA mutation and/or a first degree relative with a LRRK2, GBA or SNCA mutation 3. Willingness to undergo genetic testing, but may choose either to be informed of the results or remain unaware of the results	

Table A.2: Table detailing the inclusion and exclusion criteria for participant categories in the AMP-PD PPMI dataset. Information adapted from amp-pd.org/unified-cohorts/ppmi.

Bibliography

[1] Sabrina D. Dyall, Mark T. Brown, and Patricia J. Johnson. Ancient Invasions: From Endosymbionts to Organelles. *Science*, 304(5668):253–257, April 2004. Publisher: American Association for the Advancement of Science.

[2] Jeremy N. Timmis, Michael A. Ayliffe, Chun Y. Huang, and William Martin. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, 5(2):123–135, February 2004. Number: 2 Publisher: Nature Publishing Group.

[3] Tim R. Mercer, Shane Neph, Marcel E. Dinger, Joanna Crawford, Martin A. Smith, Anne-Marie J. Shearwood, Eric Haugen, Cameron P. Bracken, Oliver Rackham, John A. Stamatoyannopoulos, Aleksandra Filipovska, and John S. Mattick. The Human Mitochondrial Transcriptome. *Cell*, 146(4):645–658, August 2011.

[4] Y. Aloni and G. Attardi. Symmetrical In Vivo Transcription of Mitochondrial DNA in HeLa Cells. *Proceedings of the National Academy of Sciences of the United States of America*, 68(8):1757–1761, August 1971.

[5] Brendan Miller, Su-Jeong Kim, Hiroshi Kumagai, Kelvin Yen, and Pinchas Cohen. Mitochondria-derived peptides in aging and healthspan. *Journal of Clinical Investigation*, 132(9):e158449, May 2022.

[6] Xu Liu and Ge Shan. Mitochondria Encoded Non-coding RNAs in Cell Physiology. *Frontiers in Cell and Developmental Biology*, 9:713729, 2021.

[7] Gilad Barshad, Shani Marom, Tal Cohen, and Dan Mishmar. Mitochondrial DNA Transcription and Its Regulation: An Evolutionary Perspective. *Trends in Genetics*, 34(9):682–692, September 2018.

[8] H.S. Hillen, Y.I. Morozov, A. Sarfallah, D. Temiakov, and P. Cramer. Structural Basis of Mitochondrial Transcription Initiation. *Cell*, 171(5):1072.e10–1081.e10, 2017.

[9] Yaroslav I. Morozov, Karen Agaronyan, Alan C. M. Cheung, Michael Anikin, Patrick Cramer, and Dmitry Temiakov. A novel intermediate in transcription initiation by human mitochondrial RNA polymerase. *Nucleic Acids Research*, 42(6):3884–3893, April 2014.

[10] Pietro Boccaletto, Filip Stefaniak, Angana Ray, Andrea Cappannini, Sunandan Mukherjee, Elżbieta Purta, Małgorzata Kurkowska, Niloofar Shirvanizadeh, Eliana Destefanis, Paula Groza, Gülben Avşar, Antonia Romitelli, Pınar Pir, Erik Dassi, Silvestro G Conticello, Francesca Aguilo, and Janusz M Bujnicki. MODOMICS: a database of RNA modification pathways. 2021 update. *Nucleic Acids Research*, 50(D1):D231–D235, January 2022.

[11] Pedro Rebelo-Guiomar, Christopher A. Powell, Lindsey Van Haute, and Michal Minczuk. The mammalian mitochondrial epitranscriptome. *Biochimica et Biophysica Acta. Gene Regulatory Mechanisms*, 1862(3):429–446, March 2019.

[12] Markus T. Bohnsack and Katherine E. Sloan. The mitochondrial epitranscriptome: the roles of RNA modifications in mitochondrial translation and human disease. *Cellular and Molecular Life Sciences*, 75(2):241–260, January 2018.

[13] Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic RNA Modifications in Gene Expression Regulation. *Cell*, 169(7):1187–1200, June 2017.

[14] Takeo Suzuki, Yuka Yashiro, Ittoku Kikuchi, Yuma Ishigami, Hironori Saito, Ikuya Matsuzawa, Shunpei Okada, Mari Mito, Shintaro Iwasaki, Ding Ma, Xuewei Zhao, Kana Asano, Huan Lin, Yohei Kirino, Yuriko Sakaguchi, and Tsutomu Suzuki. Complete chemical structures of human mitochondrial tRNAs. *Nature Communications*, 11(1):4269, August 2020. Number: 1 Publisher: Nature Publishing Group.

[15] Deanna Ojala, Julio Montoya, and Giuseppe Attardi. tRNA punctuation model of RNA processing in human mitochondria. *Nature*, 290(5806):470–474, April 1981.

[16] Felix Voigts-Hoffmann, Martin Hengesbach, Andrei Yu Kobitski, Arthur van Aerschot, Piet Herdewijn, G. Ulrich Nienhaus, and Mark Helm. A methyl group controls conformational equilibrium in human mitochondrial tRNA(Lys). *Journal of the American Chemical Society*, 129(44):13382–13383, November 2007.

[17] M. Helm, R. Giegé, and C. Florentz. A Watson-Crick base-pair-disrupting methyl group (m1A9) is sufficient for cloverleaf folding of human mitochondrial tRNALys. *Biochemistry*, 38(40):13338–13346, October 1999.

[18] Uwe Richter, Molly E. Evans, Wesley C. Clark, Paula Marttinen, Eric A. Shoubridge, Anu Suomalainen, Anna Wredenberg, Anna Wedell, Tao Pan, and Brendan J. Battersby. RNA modification landscape of the human mitochondrial tRNALys regulates protein synthesis. *Nature Communications*, 9(1):3966, September 2018. Number: 1 Publisher: Nature Publishing Group.

[19] Youssef Idaghdour and Alan Hodgkinson. Integrated genomic analysis of mitochondrial RNA processing in human cancers. *Genome Medicine*, 9(1):36, December 2017.

[20] Talisa K. Silzer, Gita A. Pathak, and Nicole R. Phillips. Mitochondrial tRNA methylation in Alzheimer's disease and progressive supranuclear palsy. *BMC Medical Genomics*, 13(1):71, May 2020.

[21] Aminah Tasnim Ali, Youssef Idaghdour, and Alan Hodgkinson. Analysis of mitochondrial m1A/G RNA modification reveals links to nuclear genetic variants and associated disease processes. *Communications Biology*, 3(1):1–11, March 2020. Number: 1 Publisher: Nature Publishing Group.

[22] Alan Hodgkinson, Youssef Idaghdour, Elias Gbeha, Jean-Christophe Grenier, Elodie Hip-Ki, Vanessa Bruat, Jean-Philippe Goulet, Thibault de Malliard, and Philip Awadalla. High-Resolution Genomic Analysis of Human Mitochondrial RNA Sequence Variation. *Science*, 344(6182):413–415, April 2014.

[23] Roberta Filograna, Mara Mennuni, David Alsina, and Nils-Göran Larsson. Mitochondrial DNA copy number in human disease: the more the better? *Febs Letters*, 595(8):976–1002, April 2021.

[24] Martin Picard. Blood mitochondrial DNA copy number: What are we counting? *Mitochondrion*, 60:1–11, September 2021.

[25] Ren-Kui Bai and Lee-Jun C. Wong. Simultaneous detection and quantification of mitochondrial DNA deletion(s), depletion, and over-replication in patients with mitochondrial disease. *The Journal of molecular diagnostics: JMD*, 7(5):613–622, November 2005.

[26] Bruno José Moraes, Patrícia Coelho, Lígia Fão, Ildete Luísa Ferreira, and A. Cristina Rego. Modified Glutamatergic Postsynapse in Neurodegenerative Disorders. *Neuroscience*, 454:116–139, February 2021.

[27] Stephanie Y. Yang, Christina A. Castellani, Ryan J. Longchamps, Vamsee K. Pillalamarri, Brian O'Rourke, Eliseo Guallar, and Dan E. Arking. Blood-derived mitochondrial DNA copy number is associated with gene expression across multiple tissues and is predictive for incident neurodegenerative disease. *Genome Research*, 31(3):349–358, March 2021.

[28] Rahul Gupta, Masahiro Kanai, Timothy J. Durham, Kristin Tsuo, Jason G. McCoy, Patrick F. Chinnery, Konrad J. Karczewski, Sarah E. Calvo,

Benjamin M. Neale, and Vamsi K. Mootha. Nuclear genetic control of mtDNA copy number and heteroplasmy in humans, January 2023. Pages: 2023.01.19.23284696.

[29] James B. Stewart and Patrick F. Chinnery. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. *Nature Reviews Genetics*, 16(9):530–542, September 2015. Number: 9 Publisher: Nature Publishing Group.

[30] M. Sciacco, E. Bonilla, E. A. Schon, S. DiMauro, and C. T. Moraes. Distribution of wild-type and common deletion forms of mtDNA in normal and respiration-deficient muscle fibers from patients with mitochondrial myopathy. *Human Molecular Genetics*, 3(1):13–19, January 1994.

[31] N G Larsson, M H Tulinius, E Holme, A Oldfors, O Andersen, J Wahlström, and J Aasly. Segregation and manifestations of the mtDNA tRNA(Lys) A–>G(8344) mutation of myoclonus epilepsy and ragged-red fibers (MERRF) syndrome. *American Journal of Human Genetics*, 51(6):1201–1212, December 1992.

[32] E. J. Brierley, M. A. Johnson, R. N. Lightowlers, O. F. James, and D. M. Turnbull. Role of mitochondrial DNA mutations in human aging: implications for the central nervous system and muscle. *Annals of Neurology*, 43(2):217–223, February 1998.

[33] Andreas Bender, Kim J. Krishnan, Christopher M. Morris, Geoffrey A. Taylor, Amy K. Reeve, Robert H. Perry, Evelyn Jaros, Joshua S. Hersheson, Joanne Betts, Thomas Klopstock, Robert W. Taylor, and Douglass M. Turnbull. High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease. *Nature Genetics*, 38(5):515–517, May 2006.

[34] Yevgenya Kraytsberg, Elena Kudryavtseva, Ann C. McKee, Changiz Geula, Neil W. Kowall, and Konstantin Khrapko. Mitochondrial DNA deletions are

abundant and cause functional impairment in aged human substantia nigra neurons. *Nature Genetics*, 38(5):518–520, May 2006.

[35] Carlotta Giorgi, Federica Baldassari, Angela Bononi, Massimo Bonora, Elena De Marchi, Saverio Marchi, Sonia Missiroli, Simone Patergnani, Alessandro Rimessi, Jan M. Suski, Mariusz R. Wieckowski, and Paolo Pinton. Mitochondrial Ca2+ and apoptosis. *Cell Calcium*, 52(1):36–43, July 2012.

[36] Brian Billups and Ian D. Forsythe. Presynaptic mitochondrial calcium sequestration influences transmission at mammalian central synapses. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 22(14):5840–5847, July 2002.

[37] Guido Kroemer and John C. Reed. Mitochondrial control of cell death. *Nature Medicine*, 6(5):513–519, May 2000. Number: 5 Publisher: Nature Publishing Group.

[38] Yinjuan Song, Yang Zhou, and Xiangmei Zhou. The role of mitophagy in innate immune responses triggered by mitochondrial stress. *Cell Communication and Signaling*, 18(1):186, November 2020.

[39] Erol C. Bayraktar, Lou Baudrier, Ceren Özerdem, Caroline A. Lewis, Sze Ham Chan, Tenzin Kunchok, Monther Abu-Remaileh, Andrew L. Cangelosi, David M. Sabatini, Kıvanç Birsoy, and Walter W. Chen. MITO-Tag Mice enable rapid isolation and multimodal profiling of mitochondria from specific cell types in vivo. *Proceedings of the National Academy of Sciences*, 116(1):303–312, January 2019. Publisher: Proceedings of the National Academy of Sciences.

[40] Garrett M. Fogo, Anthony R. Anzell, Kathleen J. Maheras, Sarita Raghunayakula, Joseph M. Wider, Katlynn J. Emaus, Timothy D. Bryson, Melissa J. Bukowski, Robert W. Neumar, Karin Przyklenk, and Thomas H.

Sanderson. Machine learning-based classification of mitochondrial morphology in primary neurons and brain. *Scientific Reports*, 11:5133, March 2021.

[41] Caroline Fecher, Laura Trovò, Stephan A. Müller, Nicolas Snaidero, Jennifer Wettmarshausen, Sylvia Heink, Oskar Ortiz, Ingrid Wagner, Ralf Kühn, Jana Hartmann, Rosa Maria Karl, Arthur Konnerth, Thomas Korn, Wolfgang Wurst, Doron Merkler, Stefan F. Lichtenthaler, Fabiana Perocchi, and Thomas Misgeld. Cell-type-specific profiling of brain mitochondria reveals functional and molecular diversity. *Nature Neuroscience*, 22(10):1731–1742, October 2019.

[42] Chong Wang, Wanqing Du, Qian Peter Su, Mingli Zhu, Peiyuan Feng, Ying Li, Yichen Zhou, Na Mi, Yueyao Zhu, Dong Jiang, Senyan Zhang, Zerui Zhang, Yujie Sun, and Li Yu. Dynamic tubulation of mitochondria drives mitochondrial network formation. *Cell Research*, 25(10):1108–1120, October 2015. Number: 10 Publisher: Nature Publishing Group.

[43] Patries M. Herst, Matthew R. Rowe, Georgia M. Carson, and Michael V. Berridge. Functional Mitochondria in Health and Disease. *Frontiers in Endocrinology*, 8:296, November 2017.

[44] E. Gottlieb, S. M. Armour, M. H. Harris, and C. B. Thompson. Mitochondrial membrane potential regulates matrix configuration and cytochrome c release during apoptosis. *Cell Death & Differentiation*, 10(6):709–717, June 2003. Number: 6 Publisher: Nature Publishing Group.

[45] Omer Papier, Gavriel Minor, Hadar Medini, and Dan Mishmar. Coordination of mitochondrial and nuclear gene-expression regulation in health, evolution, and disease. *Current Opinion in Physiology*, 27:100554, June 2022.

[46] Aminah T Ali, Lena Boehme, Guillermo Carbajosa, Vlad C Seitan, Kerrin S Small, and Alan Hodgkinson. Nuclear genetic regulation of the human mitochondrial transcriptome. *eLife*, 8:e41927, February 2019. Publisher: eLife Sciences Publications, Ltd.

[47] Jialin C. Zheng and Shengdi Chen. Translational Neurodegeneration in the era of fast growing international brain research. *Translational Neurodegeneration*, 11(1):1, January 2022.

[48] Justin M. Long and David M. Holtzman. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell*, 179(2):312–339, October 2019.

[49] The Lancet. Lecanemab for Alzheimer's disease: tempering hype and hope. *The Lancet*, 400(10367):1899, December 2022. Publisher: Elsevier.

[50] Lecanemab in Early Alzheimer's Disease | NEJM.

[51] Francesca Alves, Pawel Kallinowski, and Scott Ayton. Accelerated Brain Volume Loss Caused by Anti-–Amyloid Drugs: A Systematic Review and Meta-analysis. *Neurology*, March 2023. Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Research Article.

[52] Valentina Escott-Price and John Hardy. Genome-wide association studies for Alzheimer's disease: bigger is not always better. *Brain Communications*, 4(3):fcac125, May 2022.

[53] Analysis of shared heritability in common disorders of the brain. *Science (New York, N.Y.)*, 360(6395):eaap8757, June 2018.

[54] Shea J. Andrews, Alan E. Renton, Brian Fulton-Howard, Anna Podlesny-Drabiniok, Edoardo Marcora, and Alison M. Goate. The complex genetic architecture of Alzheimer's disease: novel insights and future directions. *eBioMedicine*, 90, April 2023. Publisher: Elsevier.

[55] Caroline Van Cauwenberghe, Christine Van Broeckhoven, and Kristel Sleegers. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in Medicine*, 18(5):421–430, May 2016. Number: 5 Publisher: Nature Publishing Group.

[56] Douglas P. Wightman, Iris E. Jansen, Jeanne E. Savage, Alexey A. Shadrin, Shahram Bahrami, Dominic Holland, Arvid Rongve, Sigrid Børte, Bendik S. Winsvold, Ole Kristian Drange, Amy E. Martinsen, Anne Heidi Skogholt, Cristen Willer, Geir Bråthen, Ingunn Bosnes, Jonas Bille Nielsen, Lars G. Fritsche, Laurent F. Thomas, Linda M. Pedersen, Maiken E. Gabrielsen, Marianne Bakke Johnsen, Tore Wergeland Meisingset, Wei Zhou, Petroula Proitsi, Angela Hodges, Richard Dobson, Latha Velayudhan, Karl Heilbron, Adam Auton, Julia M. Sealock, Lea K. Davis, Nancy L. Pedersen, Chandra A. Reynolds, Ida K. Karlsson, Sigurdur Magnusson, Hreinn Stefansson, Steinunn Thordardottir, Palmi V. Jonsson, Jon Snaedal, Anna Zettergren, Ingmar Skoog, Silke Kern, Margda Waern, Henrik Zetterberg, Kaj Blennow, Eystein Stordal, Kristian Hveem, John-Anker Zwart, Lavinia Athanasiu, Per Selnes, Ingvild Saltvedt, Sigrid B. Sando, Ingun Ulstein, Srdjan Djurovic, Tormod Fladby, Dag Aarsland, Geir Selbæk, Stephan Ripke, Kari Stefansson, Ole A. Andreassen, and Danielle Posthuma. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nature Genetics*, 53(9):1276–1282, September 2021. Number: 9 Publisher: Nature Publishing Group.

[57] Céline Bellenguez, Fahri Küçükali, Iris E. Jansen, Luca Kleineidam, Sonia Moreno-Grau, Najaf Amin, Adam C. Naj, Rafael Campos-Martin, Benjamin Grenier-Boley, Victor Andrade, Peter A. Holmans, Anne Boland, Vincent Damotte, Sven J. van der Lee, Marcos R. Costa, Teemu Kuulasmaa, Qiong Yang, Itziar de Rojas, Joshua C. Bis, Amber Yaqub, Ivana Prokic, Julien Chapuis, Shahzad Ahmad, Vilmantas Giedraitis, Dag Aarsland, Pablo Garcia-Gonzalez, Carla Abdelnour, Emilio Alarcón-Martín, Daniel Alcolea, Montserrat Alegret, Ignacio Alvarez, Victoria Álvarez, Nicola J. Armstrong, Anthoula Tsolaki, Carmen Antúnez, Ildebrando Appollonio, Marina Arcaro, Silvana Archetti, Alfonso Arias Pastor, Beatrice Arosio, Lavinia Athanasiu, Henri Bailly, Nerisa Banaj, Miquel Baquero, Sandra Barral, Alexa Beiser, Ana Belén Pastor, Jennifer E. Below, Penelope Benchek,

Luisa Benussi, Claudine Berr, Céline Besse, Valentina Bessi, Giuliano Binetti, Alessandra Bizarro, Rafael Blesa, Mercè Boada, Eric Boerwinkle, Barbara Borroni, Silvia Boschi, Paola Bossù, Geir Bråthen, Jan Bressler, Catherine Bresner, Henry Brodaty, Keeley J. Brookes, Luis Ignacio Brusco, Dolores Buiza-Rueda, Katharina Bûrger, Vanessa Burholt, William S. Bush, Miguel Calero, Laura B. Cantwell, Geneviève Chene, Jaeyoon Chung, Michael L. Cuccaro, Ángel Carracedo, Roberta Cecchetti, Laura Cervera-Carles, Camille Charbonnier, Hung-Hsin Chen, Caterina Chillotti, Simona Ciccone, Jurgen A. H. R. Claassen, Christopher Clark, Elisa Conti, Anaïs Corma-Gómez, Emanuele Costantini, Carlo Custodero, Delphine Daian, Maria Carolina Dalmasso, Antonio Daniele, Efthimios Dardiotis, Jean-François Dartigues, Peter Paul de Deyn, Katia de Paiva Lopes, Lot D. de Witte, Stéphanie Debette, Jürgen Deckert, Teodoro del Ser, Nicola Denning, Anita DeStefano, Martin Dichgans, Janine Diehl-Schmid, Mónica Diez-Fairen, Paolo Dionigi Rossi, Srdjan Djurovic, Emmanuelle Duron, Emrah Düzel, Carole Dufouil, Gudny Eiriksdottir, Sebastiaan Engelborghs, Valentina Escott-Price, Ana Espinosa, Michael Ewers, Kelley M. Faber, Tagliavini Fabrizio, Sune Fallgaard Nielsen, David W. Fardo, Lucia Farotti, Chiara Fenoglio, Marta Fernández-Fuertes, Raffaele Ferrari, Catarina B. Ferreira, Evelyn Ferri, Bertrand Fin, Peter Fischer, Tormod Fladby, Klaus Fließbach, Bernard Fongang, Myriam Fornage, Juan Fortea, Tatiana M. Foroud, Silvia Fostinelli, Nick C. Fox, Emlio Franco-Macías, María J. Bullido, Ana Frank-García, Lutz Froelich, Brian Fulton-Howard, Daniela Galimberti, Jose Maria García-Alberca, Pablo García-González, Sebastian Garcia-Madrona, Guillermo Garcia-Ribas, Roberta Ghidoni, Ina Giegling, Giaccone Giorgio, Alison M. Goate, Oliver Goldhardt, Duber Gomez-Fonseca, Antonio González-Pérez, Caroline Graff, Giulia Grande, Emma Green, Timo Grimmer, Edna Grünblatt, Michelle Grunin, Vilmundur Gudnason, Tamar Guetta-Baranes, Annakaisa Haapasalo, Georgios Hadjigeorgiou, Jonathan L. Haines, Kara L. Hamilton-Nelson, Harald Hampel, Olivier

Hanon, John Hardy, Annette M. Hartmann, Lucrezia Hausner, Janet Harwood, Stefanie Heilmann-Heimbach, Seppo Helisalmi, Michael T. Heneka, Isabel Hernández, Martin J. Herrmann, Per Hoffmann, Clive Holmes, Henne Holstege, Raquel Huerto Vilas, Marc Hulsman, Jack Humphrey, Geert Jan Biessels, Xueqiu Jian, Charlotte Johansson, Gyungah R. Jun, Yuriko Kastumata, John Kauwe, Patrick G. Kehoe, Lena Kilander, Anne Kinhult Ståhlbom, Miia Kivipelto, Anne Koivisto, Johannes Kornhuber, Mary H. Kosmidis, Walter A. Kukull, Pavel P. Kuksa, Brian W. Kunkle, Amanda B. Kuzma, Carmen Lage, Erika J. Laukka, Lenore Launer, Alessandra Lauria, Chien-Yueh Lee, Jenni Lehtisalo, Ondrej Lerch, Alberto Lleó, William Longstreth, Oscar Lopez, Adolfo Lopez de Munain, Seth Love, Malin Löwemark, Lauren Luckcuck, Kathryn L. Lunetta, Yiyi Ma, Juan Macías, Catherine A. MacLeod, Wolfgang Maier, Francesca Mangialasche, Marco Spallazzi, Marta Marquié, Rachel Marshall, Eden R. Martin, Angel Martín Montes, Carmen Martínez Rodríguez, Carlo Masullo, Richard Mayeux, Simon Mead, Patrizia Mecocci, Miguel Medina, Alun Meggy, Shima Mehrabian, Silvia Mendoza, Manuel Menéndez-González, Pablo Mir, Susanne Moebus, Merel Mol, Laura Molina-Porcel, Laura Montrreal, Laura Morelli, Fermin Moreno, Kevin Morgan, Thomas Mosley, Markus M. Nöthen, Carolina Muchnik, Shubhabrata Mukherjee, Benedetta Nacmias, Tiia Ngandu, Gael Nicolas, Børge G. Nordestgaard, Robert Olaso, Adelina Orellana, Michela Orsini, Gemma Ortega, Alessandro Padovani, Caffarra Paolo, Goran Papenberg, Lucilla Parnetti, Florence Pasquier, Pau Pastor, Gina Peloso, Alba Pérez-Cordón, Jordi Pérez-Tur, Pierre Pericard, Oliver Peters, Yolande A. L. Pijnenburg, Juan A. Pineda, Gerard Piñol-Ripoll, Claudia Pisanu, Thomas Polak, Julius Popp, Danielle Posthuma, Josef Priller, Raquel Puerta, Olivier Quenez, Inés Quintela, Jesper Qvist Thomassen, Alberto Rábano, Innocenzo Rainero, Farid Rajabli, Inez Ramakers, Luis M. Real, Marcel J. T. Reinders, Christiane Reitz, Dolly Reyes-Dumeyer, Perry Ridge, Steffi Riedel-Heller, Peter Riederer, Natalia Roberto, Eloy Rodriguez-Rodriguez,

Arvid Rongve, Irene Rosas Allende, Maitée Rosende-Roca, Jose Luis Royo, Elisa Rubino, Dan Rujescu, María Eugenia Sáez, Paraskevi Sakka, Ingvild Saltvedt, Ángela Sanabria, María Bernal Sánchez-Arjona, Florentino Sanchez-Garcia, Pascual Sánchez Juan, Raquel Sánchez-Valle, Sigrid B. Sando, Chloé Sarnowski, Claudia L. Satizabal, Michela Scamosci, Nikolaos Scarmeas, Elio Scarpini, Philip Scheltens, Norbert Scherbaum, Martin Scherer, Matthias Schmid, Anja Schneider, Jonathan M. Schott, Geir Selbæk, Davide Seripa, Manuel Serrano, Jin Sha, Alexey A. Shadrin, Olivia Skrobot, Susan Slifer, Gijsje J. L. Snijders, Hilkka Soininen, Vincenzo Solfrizzi, Alina Solomon, Yeunjoo Song, Sandro Sorbi, Oscar Sotolongo-Grau, Gianfranco Spalletta, Annika Spottke, Alessio Squassina, Eystein Stordal, Juan Pablo Tartan, Lluís Tárraga, Niccolo Tesí, Anbupalam Thalamuthu, Tegos Thomas, Giuseppe Tosto, Latchezar Traykov, Lucio Tremolizzo, Anne Tybjærg-Hansen, Andre Uitterlinden, Abbe Ullgren, Ingun Ulstein, Sergi Valero, Otto Valladares, Christine Van Broeckhoven, Jeffery Vance, Badri N. Vardarajan, Aad van der Lugt, Jasper Van Dongen, Jeroen van Rooij, John van Swieten, Rik Vandenberghe, Frans Verhey, Jean-Sébastien Vidal, Jonathan Vogelgsang, Martin Vyhnalek, Michael Wagner, David Wallon, Li-San Wang, Ruiqi Wang, Leonie Weinhold, Jens Wiltfang, Gill Windle, Bob Woods, Mary Yannakoulia, Habil Zare, Yi Zhao, Xiaoling Zhang, Congcong Zhu, Miren Zulaica, Lindsay A. Farrer, Bruce M. Psaty, Mohsen Ghanbari, Towfique Raj, Perminder Sachdev, Karen Mather, Frank Jessen, M. Arfan Ikram, Alexandre de Mendonça, Jakub Hort, Magda Tsolaki, Margaret A. Pericak-Vance, Philippe Amouyel, Julie Williams, Ruth Frikke-Schmidt, Jordi Clarimon, Jean-François Deleuze, Giacomina Rossi, Sudha Seshadri, Ole A. Andreassen, Martin Ingelsson, Mikko Hiltunen, Kristel Sleegers, Gerard D. Schellenberg, Cornelia M. van Duijn, Rebecca Sims, Wiesje M. van der Flier, Agustín Ruiz, Alfredo Ramirez, and Jean-Charles Lambert. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics*, 54(4):412–436, April 2022. Number: 4

Publisher: Nature Publishing Group.

[58] Russell H. Swerdlow. Mitochondria and Mitochondrial Cascades in Alzheimer's Disease. *Journal of Alzheimer's Disease*, 62(3):1403–1416, March 2018.

[59] Werner Poewe, Klaus Seppi, Caroline M. Tanner, Glenda M. Halliday, Patrik Brundin, Jens Volkmann, Anette-Eleonore Schrag, and Anthony E. Lang. Parkinson disease. *Nature Reviews Disease Primers*, 3(1):1–21, March 2017. Number: 1 Publisher: Nature Publishing Group.

[60] Claire McDonald, Gavin Gordon, Annette Hand, Richard W Walker, and James M Fisher. 200 Years of Parkinson's disease: what have we learnt from James Parkinson? *Age and Ageing*, 47(2):209–214, March 2018.

[61] Cornelis Blauwendraat, Mike A Nalls, and Andrew B Singleton. The genetic architecture of Parkinson's disease. *The Lancet Neurology*, 19(2):170–178, February 2020.

[62] Nicola Tambasco, Michele Romoli, and Paolo Calabresi. Levodopa in Parkinson's Disease: Current Status and Future Developments. *Current Neuropharmacology*, 16(8):1239, October 2018. Publisher: Bentham Science Publishers.

[63] Mike A. Nalls, Valentina Escott-Price, Nigel M. Williams, Steven Lubbe, Margaux F. Keller, Huw R. Morris, Andrew B. Singleton, and on behalf of the International Parkinson's Disease Genomics Consortium (IPDGC). Genetic risk and age in Parkinson's disease: Continuum not stratum. *Movement Disorders*, 30(6):850–854, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.26192.

[64] Mihael H. Polymeropoulos, Christian Lavedan, Elisabeth Leroy, Susan E. Ide, Anindya Dehejia, Amalia Dutra, Brian Pike, Holly Root, Jeffrey Rubenstein, Rebecca Boyer, Edward S. Stenroos, Settara Chandrasekharappa,

Aglaia Athanassiadou, Theodore Papapetropoulos, William G. Johnson, Alice M. Lazzarini, Roger C. Duvoisin, Giuseppe Di Iorio, Lawrence I. Golbe, and Robert L. Nussbaum. Mutation in the -Synuclein Gene Identified in Families with Parkinson's Disease. *Science*, 276(5321):2045–2047, June 1997.

[65] Mike A Nalls, Cornelis Blauwendraat, Costanza L Vallerga, Karl Heilbron, Sara Bandres-Ciga, Diana Chang, Manuela Tan, Demis A Kia, Alastair J Noyce, Angli Xue, Jose Bras, Emily Young, Rainer von Coelln, Javier Simón-Sánchez, Claudia Schulte, Manu Sharma, Lynne Krohn, Lasse Pihlstrøm, Ari Siitonen, Hirotaka Iwaki, Hampton Leonard, Faraz Faghri, J Raphael Gibbs, Dena G Hernandez, Sonja W Scholz, Juan A Botia, Maria Martinez, Jean-Christophe Corvol, Suzanne Lesage, Joseph Jankovic, Lisa M Shulman, Margaret Sutherland, Pentti Tienari, Kari Majamaa, Mathias Toft, Ole A Andreassen, Tushar Bangale, Alexis Brice, Jian Yang, Ziv Gan-Or, Thomas Gasser, Peter Heutink, Joshua M Shulman, Nicholas W Wood, David A Hinds, John A Hardy, Huw R Morris, Jacob Gratten, Peter M Visscher, Robert R Graham, Andrew B Singleton, Astrid D Adarmes-Gómez, Miquel Aguilar, Akbota Aitkulova, Vadim Akhmetzhanov, Roy N Alcalay, Ignacio Alvarez, Victoria Alvarez, Sara Bandres-Ciga, Francisco Javier Barrero, Jesús Alberto Bergareche Yarza, Inmaculada Bernal-Bernal, Kimberley Billingsley, Cornelis Blauwendraat, Marta Blazquez, Marta Bonilla-Toribio, Juan A Botía, María Teresa Boungiorno, Jose Bras, Alexis Brice, Kathrin Brockmann, Vivien Bubb, Dolores Buiza-Rueda, Ana Cámara, Fátima Carrillo, Mario Carrión-Claro, Debora Cerdan, Viorica Chelban, Jordi Clarimón, Carl Clarke, Yaroslau Compta, Mark R Cookson, Jean-Christophe Corvol, David W Craig, Fabrice Danjou, Monica Diez-Fairen, Oriol Dols-Icardo, Jacinto Duarte, Raquel Duran, Francisco Escamilla-Sevilla, Valentina Escott-Price, Mario Ezquerra, Faraz Faghri, Cici Feliz, Manel Fernández, Rubén Fernández-Santiago, Steven Finkbeiner, Thomas Foltynie, Ziv Gan-Or, Ciara Garcia, Pedro García-Ruiz, Thomas Gasser, J Raphael Gibbs, Maria Jose Gomez Heredia, Pilar Gómez-Garre, Manuel Menéndez González, Isabel

Gonzalez-Aramburu, Sebastian Guelfi, Rita Guerreiro, John Hardy, Sharon Hassin-Baer, Dena G Hernandez, Peter Heutink, Janet Hoenicka, Peter Holmans, Henry Houlden, Jon Infante, Hirotaka Iwaki, Silvia Jesús, Adriano Jimenez-Escrig, Gulnaz Kaishybayeva, Rauan Kaiyrzhanov, Altynay Karimova, Demis A Kia, Kerri J Kinghorn, Sulev Koks, Lynne Krohn, Jaime Kulisevsky, Miguel A Labrador-Espinosa, Hampton L Leonard, Suzanne Lesage, Patrick Lewis, Jose Luis Lopez-Sendon, Ruth Lovering, Steven Lubbe, Codrin Lungu, Daniel Macias, Kari Majamaa, Claudia Manzoni, Juan Marín, Johan Marinus, Maria Jose Marti, Maria Martinez, Irene Martínez Torres, Juan Carlos Martínez-Castrillo, Marina Mata, Niccolo E Mencacci, Carlota Méndez-del Barrio, Ben Middlehurst, Adolfo Mínguez, Pablo Mir, Kin Y Mok, Huw R Morris, Esteban Muñoz, Mike A Nalls, Derek Narendra, Alastair J Noyce, Oluwadamilola O Ojo, Njideka U Okubadejo, Ana Gorostidi Pagola, Pau Pastor, Francisco Perez Errazquin, Teresa Periñán-Tocino, Lasse Pihlstrom, Helene Plun-Favreau, John Quinn, Lea R'Bibo, Xylena Reed, Elisabet Mondragon Rezola, Mie Rizig, Patrizia Rizzu, Laurie Robak, Antonio Sanchez Rodriguez, Guy A Rouleau, Javier Ruiz-Martínez, Clara Ruz, Mina Ryten, Dinara Sadykova, Sonja W Scholz, Sebastian Schreglmann, Claudia Schulte, Manu Sharma, Chingiz Shashkin, Joshua M Shulman, María Sierra, Ari Siitonen, Javier Simón-Sánchez, Andrew B Singleton, Esther Suarez-Sanmartin, Pille Taba, Cesar Tabernero, Manuela X Tan, Juan Pablo Tartari, Cristina Tejera-Parrado, Mathias Toft, Eduard Tolosa, Daniah Trabzuni, Francesc Valldeoriola, Jacobus J van Hilten, Kendall Van Keuren-Jensen, Laura Vargas-González, Lydia Vela, Francisco Vives, Nigel Williams, Nicholas W Wood, Nazira Zharkinbekova, Zharkyn Zharmukhanov, Elena Zholdybayeva, Alexander Zimprich, Pauli Ylikotila, Lisa M. Shulman, Rainer von Coelln, Stephen Reich, Joseph Savitt, Michelle Agee, Babak Alipanahi, Adam Auton, Robert K. Bell, Katarzyna Bryc, Sarah L. Elson, Pierre Fontanillas, Nicholas A. Furlotte, Karen E. Huber, Barry Hicks, Ethan M. Jewett, Yunxuan Jiang, Aaron

Kleinman, Keng-Han Lin, Nadia K. Litterman, Jennifer C. McCreight, Matthew H. McIntyre, Kimberly F. McManus, Joanna L. Mountain, Elizabeth S. Noblin, Carrie A.M. Northover, Steven J. Pitts, G. David Poznik, J. Fah Sathirapongsasuti, Janie F. Shelton, Suyash Shringarpure, Chao Tian, Joyce Tung, Vladimir Vacic, Xin Wang, Catherine H. Wilson, Tim Anderson, Steven Bentley, John Dalrymple-Alford, Javed Fowdar, Jacob Gratten, Glenda Halliday, Anjali K. Henders, Ian Hickie, Irfahan Kassam, Martin Kennedy, John Kwok, Simon Lewis, George Mellick, Grant Montgomery, John Pearson, Toni Pitcher, Julia Sidorenko, Peter A. Silburn, Costanza L. Vallerga, Peter M. Visscher, Leanne Wallace, Naomi R. Wray, Angli Xue, Jian Yang, and Futao Zhang. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *The Lancet Neurology*, 18(12):1091–1102, December 2019.

[66] Emil Uffelmann, Qin Qin Huang, Nchangwi Syntia Munung, Jantina de Vries, Yukinori Okada, Alicia R. Martin, Hilary C. Martin, Tuuli Lappalainen, and Danielle Posthuma. Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21, August 2021. Number: 1 Publisher: Nature Publishing Group.

[67] J. William Langston, Philip Ballard, James W. Tetrud, and Ian Irwin. Chronic Parkinsonism in Humans Due to a Product of Meperidine-Analog Synthesis. *Science*, 219(4587):979–980, February 1983. Publisher: American Association for the Advancement of Science.

[68] Anindita Bose and M. Flint Beal. Mitochondrial dysfunction in Parkinson's disease. *Journal of Neurochemistry*, 139(S1):216–231, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jnc.13731.

[69] Latha Devi, Vijayendran Raghavendran, Badanavalu M. Prabhu, Narayan G. Avadhani, and Hindupur K. Anandatheerthavarada. Mitochondrial Import and Accumulation of -Synuclein Impair Complex I in Human Dopaminer-

gic Neuronal Cultures and Parkinson Disease Brain. *Journal of Biological Chemistry*, 283(14):9089–9100, April 2008.

[70] Mats I. Ekstrand, Mügen Terzioglu, Dagmar Galter, Shunwei Zhu, Christoph Hofstetter, Eva Lindqvist, Sebastian Thams, Anita Bergstrand, Fredrik Sterky Hansson, Aleksandra Trifunovic, Barry Hoffer, Staffan Cullheim, Abdul H. Mohammed, Lars Olson, and Nils-Göran Larsson. Progressive parkinsonism in mice with respiratory-chain-deficient dopamine neurons. *Proceedings of the National Academy of Sciences*, 104(4):1325–1330, January 2007.

[71] D. F. Rolfe and G. C. Brown. Cellular energy utilization and molecular origin of standard metabolic rate in mammals. *Physiological Reviews*, 77(3):731–758, July 1997.

[72] Eleftheria K. Pissadaki and J. Paul Bolam. The energy cost of action potential propagation in dopamine neurons: clues to susceptibility in Parkinson's disease. *Frontiers in Computational Neuroscience*, 7, 2013.

[73] Fahmeed Hyder, Douglas L. Rothman, and Maxwell R. Bennett. Cortical energy demands of signaling and nonsignaling components in brain are conserved across mammalian species and activity levels. *Proceedings of the National Academy of Sciences*, 110(9):3549–3554, February 2013.

[74] Chris Meisinger, Albert Sickmann, and Nikolaus Pfanner. The Mitochondrial Proteome: From Inventory to Function. *Cell*, 134(1):22–24, July 2008.

[75] Sneha Rath, Rohit Sharma, Rahul Gupta, Tslil Ast, Connie Chan, Timothy J Durham, Russell P Goodman, Zenon Grabarek, Mary E Haas, Wendy H W Hung, Pallavi R Joshi, Alexis A Jourdain, Sharon H Kim, Anna V Kotrys, Stephanie S Lam, Jason G McCoy, Joshua D Meisel, Maria Miranda, Apekshya Panda, Anupam Patgiri, Robert Rogers, Shayan Sadre, Hardik Shah, Owen S Skinner, Tsz-Leung To, Melissa A Walker, Hong Wang, Patrick S Ward, Jordan Wengrod, Chen-Ching Yuan, Sarah E Calvo, and

Vamsi K Mootha. MitoCarta3.0: an updated mitochondrial proteome now with sub-organelle localization and pathway annotations. *Nucleic Acids Research*, 49(D1):D1541–D1547, January 2021.

[76] Nikolaus Pfanner, Bettina Warscheid, and Nils Wiedemann. Mitochondrial proteins: from biogenesis to functional networks. *Nature Reviews Molecular Cell Biology*, 20(5):267–284, May 2019.

[77] David Attwell and Simon B. Laughlin. An Energy Budget for Signaling in the Grey Matter of the Brain. *Journal of Cerebral Blood Flow & Metabolism*, 21(10):1133–1145, October 2001. Publisher: SAGE Publications Ltd STM.

[78] A. J. Hulbert, Reinald Pamplona, Rochelle Buffenstein, and W. A. Buttemer. Life and Death: Metabolic Rate, Membrane Composition, and Life Span of Animals. *Physiological Reviews*, 87(4):1175–1213, October 2007. Publisher: American Physiological Society.

[79] Gerben van Hameren, Graham Campbell, Marie Deck, Jade Berthelot, Benoit Gautier, Patrice Quintana, Roman Chrast, and Nicolas Tricaud. In vivo real-time dynamics of ATP and ROS production in axonal mitochondria show decoupling in mouse models of peripheral neuropathies. *Acta Neuropathologica Communications*, 7(1):86, June 2019.

[80] Susana Cadenas. Mitochondrial uncoupling, ROS generation and cardioprotection. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1859(9):940–950, September 2018.

[81] Anatoly A. Starkov. The Role of Mitochondria in Reactive Oxygen Species Metabolism and Signaling. *Annals of the New York Academy of Sciences*, 1147(1):37–52, December 2008.

[82] Mariusz Karbowski and Albert Neutzner. Neurodegeneration as a consequence of failed mitochondrial maintenance. *Acta Neuropathologica*, 123(2):157–171, February 2012.

[83] Baris Bingol and Morgan Sheng. Deconstruction for Reconstruction: The Role of Proteolysis in Neural Plasticity and Disease. *Neuron*, 69(1):22–32, January 2011.

[84] Noriyuki Matsuda, Shigeto Sato, Kahori Shiba, Kei Okatsu, Keiko Saisho, Clement A. Gautier, Yu-shin Sou, Shinji Saiki, Sumihiro Kawajiri, Fumiaki Sato, Mayumi Kimura, Masaaki Komatsu, Nobutaka Hattori, and Keiji Tanaka. PINK1 stabilized by mitochondrial depolarization recruits Parkin to damaged mitochondria and activates latent Parkin for mitophagy. *Journal of Cell Biology*, 189(2):211–221, April 2010.

[85] Derek P. Narendra, Seok Min Jin, Atsushi Tanaka, Der-Fen Suen, Clement A. Gautier, Jie Shen, Mark R. Cookson, and Richard J. Youle. PINK1 Is Selectively Stabilized on Impaired Mitochondria to Activate Parkin. *PLoS Biology*, 8(1):e1000298, January 2010.

[86] T. L. Schwarz. Mitochondrial Trafficking in Neurons. *Cold Spring Harbor Perspectives in Biology*, 5(6):a011304–a011304, June 2013.

[87] Meredith M. Course and Xinnan Wang. Transporting mitochondria in neurons. *F1000Research*, 5:1735, July 2016.

[88] D. T. W. Chang, A. S. Honick, and I. J. Reynolds. Mitochondrial Trafficking to Synapses in Cultured Primary Cortical Neurons. *Journal of Neuroscience*, 26(26):7035–7045, June 2006.

[89] Christine E. Holt, Kelsey C. Martin, and Erin M. Schuman. Local translation in neurons: visualization and function. *Nature Structural & Molecular Biology*, 26(7):557–566, July 2019.

[90] Jeehye Park, Sung Bae Lee, Sungkyu Lee, Yongsung Kim, Saera Song, Sunhong Kim, Eunkyung Bae, Jaeseob Kim, Minho Shong, Jin-Man Kim, and Jongkyeong Chung. Mitochondrial dysfunction in Drosophila PINK1 mutants is complemented by parkin. *Nature*, 441(7097):1157–1161, June 2006. Number: 7097 Publisher: Nature Publishing Group.

[91] T. Kitada, S. Asakawa, N. Hattori, H. Matsumine, Y. Yamamura, S. Minoshima, M. Yokochi, Y. Mizuno, and N. Shimizu. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*, 392(6676):605–608, April 1998.

[92] Baiyang Sheng, Xinglong Wang, Bo Su, Hyoung-gon Lee, Gemma Casadesus, George Perry, and Xiongwei Zhu. Impaired mitochondrial biogenesis contributes to mitochondrial dysfunction in Alzheimer's disease: Impaired mitochondrial biogenesis in AD. *Journal of Neurochemistry*, 120(3):419–429, February 2012.

[93] Xuan Liu, Nares Trakooljul, Frieder Hadlich, Eduard Murani, Klaus Wimmers, and Siriluck Ponsuksili. Mitochondrial-nuclear crosstalk, haplotype and copy number variation distinct in muscle fiber type, mitochondrial respiratory and metabolic enzyme activities. *Scientific Reports*, 7(1):14024, December 2017.

[94] Angelika Parl, Sabrina L. Mitchell, Hayley B. Clay, Sara Reiss, Zhen Li, and Deborah G. Murdock. The mitochondrial fatty acid synthesis (mtFASII) pathway is capable of mediating nuclear-mitochondrial cross talk through the PPAR system of transcriptional activation. *Biochemical and Biophysical Research Communications*, 441(2):418–424, November 2013.

[95] M Goldenthal, R Ananthakrishnan, and J Maringarcia. Nuclear-mitochondrial cross-talk in cardiomyocyte T3 signaling: A time-course analysis. *Journal of Molecular and Cellular Cardiology*, 39(2):319–326, August 2005.

[96] Gilad Barshad, Amit Blumberg, Tal Cohen, and Dan Mishmar. Human primitive brain displays negative mitochondrial-nuclear expression correlation of respiratory genes. *Genome Research*, 28(7):952–967, July 2018. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Lab-

oratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[97] Peter Langfelder and Steve Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1):559, December 2008.

[98] Qi Liao, Changning Liu, Xiongying Yuan, Shuli Kang, Ruoyu Miao, Hui Xiao, Guoguang Zhao, Haitao Luo, Dechao Bu, Haitao Zhao, Geir Skogerbø, Zhongdao Wu, and Yi Zhao. Large-scale prediction of long non-coding RNA functions in a coding–non-coding gene co-expression network. *Nucleic Acids Research*, 39(9):3864–3878, May 2011.

[99] Tongxin Wang, Jie Zhang, and Kun Huang. Generalized gene co-expression analysis via subspace clustering using low-rank representation. *BMC Bioinformatics*, 20(7):196, May 2019.

[100] The GTEx Consortium, Kristin G. Ardlie, David S. Deluca, Ayellet V. Segrè, Timothy J. Sullivan, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, Julian B. Maller, Taru Tukiainen, Monkol Lek, Lucas D. Ward, Pouya Kheradpour, Benjamin Iriarte, Yan Meng, Cameron D. Palmer, Tõnu Esko, Wendy Winckler, Joel N. Hirschhorn, Manolis Kellis, Daniel G. MacArthur, Gad Getz, Andrey A. Shabalin, Gen Li, Yi-Hui Zhou, Andrew B. Nobel, Ivan Rusyn, Fred A. Wright, Tuuli Lappalainen, Pedro G. Ferreira, Halit Ongen, Manuel A. Rivas, Alexis Battle, Sara Mostafavi, Jean Monlong, Michael Sammeth, Marta Mele, Ferran Reverter, Jakob M. Goldmann, Daphne Koller, Roderic Guigó, Mark I. McCarthy, Emmanouil T. Dermitzakis, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Dan L. Nicolae, Nancy J. Cox, Timothée Flutre, Xiaoquan Wen, Matthew Stephens, Jonathan K. Pritchard, Zhidong Tu, Bin Zhang, Tao Huang, Quan Long, Luan Lin, Jialiang Yang, Jun Zhu, Jun Liu, Amanda Brown, Bernadette Mestichelli, Denee Tidwell, Edmund Lo, Mike Salvatore, Saboor Shad, Jeffrey A. Thomas, John T. Lonsdale, Michael T. Moser, Bryan M. Gillard,

Ellen Karasik, Kimberly Ramsey, Christopher Choi, Barbara A. Foster, John Syron, Johnell Fleming, Harold Magazine, Rick Hasz, Gary D. Walters, Jason P. Bridge, Mark Miklos, Susan Sullivan, Laura K. Barker, Heather M. Traino, Maghboeba Mosavel, Laura A. Siminoff, Dana R. Valley, Daniel C. Rohrer, Scott D. Jewell, Philip A. Branton, Leslie H. Sobin, Mary Barcus, Liqun Qi, Jeffrey McLean, Pushpa Hariharan, Ki Sung Um, Shenpei Wu, David Tabor, Charles Shive, Anna M. Smith, Stephen A. Buia, Anita H. Undale, Karna L. Robinson, Nancy Roche, Kimberly M. Valentino, Angela Britton, Robin Burges, Debra Bradbury, Kenneth W. Hambright, John Seleski, Greg E. Korzeniewski, Kenyon Erickson, Yvonne Marcus, Jorge Tejada, Mehran Taherian, Chunrong Lu, Margaret Basile, Deborah C. Mash, Simona Volpi, Jeffery P. Struewing, Gary F. Temple, Joy Boyer, Deborah Colantuoni, Roger Little, Susan Koester, Latarsha J. Carithers, Helen M. Moore, Ping Guan, Carolyn Compton, Sherilyn J. Sawyer, Joanne P. Demchok, Jimmie B. Vaught, Chana A. Rabiner, Nicole C. Lockhart, Kristin G. Ardlie, Gad Getz, Fred A. Wright, Manolis Kellis, Simona Volpi, and Emmanouil T. Dermitzakis. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, May 2015.

[101] GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017.

[102] Philip L. De Jager, Yiyi Ma, Cristin McCabe, Jishu Xu, Badri N. Vardarajan, Daniel Felsky, Hans-Ulrich Klein, Charles C. White, Mette A. Peters, Ben Lodgson, Parham Nejad, Anna Tang, Lara M. Mangravite, Lei Yu, Chris Gaiteri, Sara Mostafavi, Julie A. Schneider, and David A. Bennett. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. *Scientific Data*, 5(1):180142, December 2018.

[103] Uku Raudvere, Liis Kolberg, Ivan Kuzmin, Tambet Arak, Priit Adler, Hedi Peterson, and Jaak Vilo. g:Profiler: a web server for functional enrichment

analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, July 2019.

[104] Frank Koopmans, Pim van Nierop, Maria Andres-Alonso, Andrea Byrnes, Tony Cijsouw, Marcelo P. Coba, L. Niels Cornelisse, Ryan J. Farrell, Hana L. Goldschmidt, Daniel P. Howrigan, Natasha K. Hussain, Cordelia Imig, Arthur P.H. de Jong, Hwajin Jung, Mahdokht Kohansalnodehi, Barbara Kramarz, Noa Lipstein, Ruth C. Lovering, Harold MacGillavry, Vittoria Mariano, Huaiyu Mi, Momchil Ninov, David Osumi-Sutherland, Rainer Pielot, Karl-Heinz Smalla, Haiming Tang, Katherine Tashman, Ruud F.G. Toonen, Chiara Verpelli, Rita Reig-Viader, Kyoko Watanabe, Jan van Weering, Tilmann Achsel, Ghazaleh Ashrafi, Nimra Asi, Tyler C. Brown, Pietro De Camilli, Marc Feuermann, Rebecca E. Foulger, Pascale Gaudet, Anoushka Joglekar, Alexandros Kanellopoulos, Robert Malenka, Roger A. Nicoll, Camila Pulido, Jaime de Juan-Sanz, Morgan Sheng, Thomas C. Südhof, Hagen U. Tilgner, Claudia Bagni, Àlex Bayés, Thomas Biederer, Nils Brose, John Jia En Chua, Daniela C. Dieterich, Eckart D. Gundelfinger, Casper Hoogenraad, Richard L. Huganir, Reinhard Jahn, Pascal S. Kaeser, Eunjoon Kim, Michael R. Kreutz, Peter S. McPherson, Ben M. Neale, Vincent O'Connor, Danielle Posthuma, Timothy A. Ryan, Carlo Sala, Guoping Feng, Steven E. Hyman, Paul D. Thomas, August B. Smit, and Matthijs Verhage. SynGO: An Evidence-Based, Expert-Curated Knowledge Base for the Synapse. *Neuron*, 103(2):217–234.e4, July 2019.

[105] Nathan G. Skene and Seth G. N. Grant. Identification of Vulnerable Cell Types in Major Brain Disorders Using Single Cell Transcriptomes and Expression Weighted Cell Type Enrichment. *Frontiers in Neuroscience*, 10, January 2016.

[106] Susan M. Sunkin, Lydia Ng, Chris Lau, Tim Dolbeare, Terri L. Gilbert, Carol L. Thompson, Michael Hawrylycz, and Chinh Dang. Allen Brain At-

las: an integrated spatio-temporal portal for exploring the central nervous system. *Nucleic Acids Research*, 41(D1):D996–D1008, November 2012.

[107] Margaret K. R. Donovan, Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, and Kelly A. Frazer. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nature Communications*, 11(1):955, December 2020.

[108] Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature Protocols*, 7(3):500–507, March 2012.

[109] Iris E. Jansen, Jeanne E. Savage, Kyoko Watanabe, Julien Bryois, Dylan M. Williams, Stacy Steinberg, Julia Sealock, Ida K. Karlsson, Sara Hägg, Lavinia Athanasiu, Nicola Voyle, Petroula Proitsi, Aree Witoelar, Sven Stringer, Dag Aarsland, Ina S. Almdahl, Fred Andersen, Sverre Bergh, Francesco Bettella, Sigurbjorn Bjornsson, Anne Brækhus, Geir Bråthen, Christiaan de Leeuw, Rahul S. Desikan, Srdjan Djurovic, Logan Dumitrescu, Tormod Fladby, Timothy J. Hohman, Palmi V. Jonsson, Steven J. Kiddle, Arvid Rongve, Ingvild Saltvedt, Sigrid B. Sando, Geir Selbæk, Maryam Shoai, Nathan G. Skene, Jon Snaedal, Eystein Stordal, Ingun D. Ulstein, Yunpeng Wang, Linda R. White, John Hardy, Jens Hjerling-Leffler, Patrick F. Sullivan, Wiesje M. van der Flier, Richard Dobson, Lea K. Davis, Hreinn Stefansson, Kari Stefansson, Nancy L. Pedersen, Stephan Ripke, Ole A. Andreassen, and Danielle Posthuma. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics*, 51(3):404–413, March 2019.

[110] Antonio Rueda Martin, Eleanor Williams, Rebecca E. Foulger, Sarah Leigh, Louise C. Daugherty, Olivia Niblock, Ivone U. S. Leong, Katherine R. Smith, Oleg Gerasimenko, Eik Haraldsdottir, Ellen Thomas, Richard H. Scott, Emma Baple, Arianna Tucci, Helen Brittain, Anna de Burca, Kristina Ibañez,

Dalia Kasperaviciute, Damian Smedley, Mark Caulfield, Augusto Rendon, and Ellen M. McDonagh. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nature Genetics*, 51(11):1560–1565, November 2019.

[111] Alexandra Grubman, Gabriel Chew, John F. Ouyang, Guizhi Sun, Xin Yi Choo, Catriona McLean, Rebecca K. Simmons, Sam Buckberry, Dulce B. Vargas-Landin, Daniel Poppe, Jahnvi Pflueger, Ryan Lister, Owen J. L. Rackham, Enrico Petretto, and Jose M. Polo. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nature Neuroscience*, 22(12):2087–2097, December 2019.

[112] Hansruedi Mathys, Jose Davila-Velderrain, Zhuyu Peng, Fan Gao, Shahin Mohammadi, Jennie Z. Young, Madhvi Menon, Liang He, Fatema Abdurrob, Xueqiao Jiang, Anthony J. Martorell, Richard M. Ransohoff, Brian P. Hafler, David A. Bennett, Manolis Kellis, and Li-Huei Tsai. Single-cell transcriptomic analysis of Alzheimer's disease. *Nature*, 570(7761):332–337, June 2019.

[113] Kevin Menden, Mohamed Marouf, Sergio Oller, Anupriya Dalmia, Daniel Sumner Magruder, Karin Kloiber, Peter Heutink, and Stefan Bonn. Deep learning–based cell composition analysis from tissue expression profiles. *Science Advances*, 6(30):eaba2619, July 2020.

[114] Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. Fast gene set enrichment analysis. preprint, Bioinformatics, June 2016.

[115] Michael Hawrylycz, Jeremy A Miller, Vilas Menon, David Feng, Tim Dolbeare, Angela L Guillozet-Bongaarts, Anil G Jegga, Bruce J Aronow, Chang-Kyu Lee, Amy Bernard, Matthew F Glasser, Donna L Dierker, Jörg Menche, Aaron Szafer, Forrest Collman, Pascal Grange, Kenneth A Berman, Stefan

Mihalas, Zizhen Yao, Lance Stewart, Albert-László Barabási, Jay Schulkin, John Phillips, Lydia Ng, Chinh Dang, David R Haynor, Allan Jones, David C Van Essen, Christof Koch, and Ed Lein. Canonical genetic signatures of the adult human brain. *Nature Neuroscience*, 18(12):1832–1844, December 2015.

[116] Naomi Habib, Inbal Avraham-Davidi, Anindita Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Sourav R Choudhury, François Aguet, Ellen Gelfand, Kristin Ardlie, David A Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nature Methods*, 14(10):955–958, October 2017.

[117] Aine Fairbrother-Browne, Aminah T. Ali, Regina H. Reynolds, Sonia Garcia-Ruiz, David Zhang, Zhongbo Chen, Mina Ryten, and Alan Hodgkinson. Mitochondrial-nuclear cross-talk in the human brain is modulated by cell type and perturbed in neurodegenerative disease. *Communications Biology*, 4(1):1262, December 2021.

[118] Ruilin Tian, Anthony Abarientos, Jason Hong, Sayed Hadi Hashemi, Rui Yan, Nina Dräger, Kun Leng, Mike A. Nalls, Andrew B. Singleton, Ke Xu, Faraz Faghri, and Martin Kampmann. Genome-wide CRISPRi/a screens in human neurons link lysosomal failure to ferroptosis. *Nature Neuroscience*, 24(7):1020–1034, July 2021.

[119] Colleen S. Stein, Pooja Jadiya, Xiaoming Zhang, Jared M. McLendon, Gabrielle M. Abouassaly, Nathan H. Witmer, Ethan J. Anderson, John W. Elrod, and Ryan L. Boudreau. Mitoregulin: A lncRNA-Encoded Microprotein that Supports Mitochondrial Supercomplexes and Respiratory Efficiency. *Cell Reports*, 23(13):3710–3720.e8, June 2018.

[120] Anastasia Chugunova, Elizaveta Loseva, Pavel Mazin, Aleksandra Mitina, Tsimafei Navalayeu, Dmitry Bilan, Polina Vishnyakova, Maria Marey, Anna Golovina, Marina Serebryakova, Philipp Pletnev, Maria Rubtsova, Waltraud

Mair, Anna Vanyushkina, Philipp Khaitovich, Vsevolod Belousov, Mikhail Vysokikh, Petr Sergiev, and Olga Dontsova. *LINC00116* codes for a mitochondrial peptide linking respiration and lipid metabolism. *Proceedings of the National Academy of Sciences*, 116(11):4940–4945, March 2019.

[121] Anarmaa Mendsaikhan, Ikuo Tooyama, Jean-Pierre Bellier, Geidy E. Serrano, Lucia I. Sue, Lih-Fen Lue, Thomas G. Beach, and Douglas G. Walker. Characterization of lysosomal proteins Progranulin and Prosaposin and their interactions in Alzheimer's disease and aged brains: increased levels correlate with neuropathology. *Acta Neuropathologica Communications*, 7(1):215, December 2019.

[122] Annika Andersson, Julia Remnestål, Bengt Nellgård, Helian Vunk, David Kotol, Fredrik Edfors, Mathias Uhlén, Jochen M. Schwenk, Leopold L. Ilag, Henrik Zetterberg, Kaj Blennow, Anna Månberg, Peter Nilsson, and Claudia Fredolini. Development of parallel reaction monitoring assays for cerebrospinal fluid proteins associated with Alzheimer's disease. *Clinica Chimica Acta*, 494:79–93, July 2019.

[123] Dennis A. Turner and David Cory Adamson. Neuronal-Astrocyte Metabolic Interactions: Understanding the Transition Into Abnormal Astrocytoma Metabolism. *Journal of Neuropathology & Experimental Neurology*, 70(3):167–176, March 2011.

[124] Mikael Simons and Klaus-Armin Nave. Oligodendrocytes: Myelination and Axonal Support. *Cold Spring Harbor Perspectives in Biology*, 8(1):a020479, January 2016.

[125] Meghan J Rossi and Gulcin Pekkurnaz. Powerhouse of the mind: mitochondrial plasticity at the synapse. *Current Opinion in Neurobiology*, 57:149–155, August 2019.

[126] Ramesh Chandra, Cali A. Calarco, and Mary Kay Lobo. Differential mito-

chondrial morphology in ventral striatal projection neuron subtypes. *Journal of Neuroscience Research*, page jnr.24511, August 2019.

[127] Csaba Erö, Marc-Oliver Gewaltig, Daniel Keller, and Henry Markram. A Cell Atlas for the Mouse Brain. *Frontiers in Neuroinformatics*, 12:84, November 2018.

[128] Blue B. Lake, Rizi Ai, Gwendolyn E. Kaeser, Neeraj S. Salathia, Yun C. Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. *Science*, 352(6293):1586–1590, June 2016.

[129] Rachel J Smith, Mary Kay Lobo, Sade Spencer, and Peter W Kalivas. Cocaine-induced adaptations in D1 and D2 accumbens projection neurons (a dichotomy not necessarily synonymous with direct and indirect pathways). *Current Opinion in Neurobiology*, 23(4):546–552, August 2013.

[130] Vyara Todorova and Arjan Blokland. Mitochondria and Synaptic Plasticity in the Mature and Aging Nervous System. *Current Neuropharmacology*, 15(1):166–173, December 2016.

[131] Scott E. Counts, Melissa J. Alldred, Shaoli Che, Stephen D. Ginsberg, and Elliott J. Mufson. Synaptic gene dysregulation within hippocampal CA1 pyramidal neurons in mild cognitive impairment. *Neuropharmacology*, 79:172–179, April 2014.

[132] Diana Marcela Cuestas Torres and Fernando P. Cardenas. Synaptic plasticity in Alzheimer's disease and healthy aging. *Reviews in the Neurosciences*, 31(3):245–268, April 2020.

[133] Lauren S. Whyte, Adeline A. Lau, Kim M. Hemsley, John J. Hopwood, and Timothy J. Sargeant. Endo-lysosomal and autophagic dysfunction: a driving

factor in Alzheimer's disease? *Journal of Neurochemistry*, 140(5):703–717, March 2017.

[134] Stjepko Cermak, Marko Kosicek, Aleksandra Mladenovic-Djordjevic, Kosara Smiljanic, Selma Kanazir, and Silva Hecimovic. Loss of Cathepsin B and L Leads to Lysosomal Dysfunction, NPC-Like Cholesterol Sequestration and Accumulation of the Key Alzheimer's Proteins. *PLOS ONE*, 11(11):e0167428, November 2016.

[135] Wesley Peng, Georgia Minakaki, Maria Nguyen, and Dimitri Krainc. Preserving Lysosomal Function in the Aging Brain: Insights from Neurodegeneration. *Neurotherapeutics*, 16(3):611–634, July 2019.

[136] Mu Yang, Stuart Matan-Lithwick, Yanling Wang, Philip L De Jager, David A Bennett, and Daniel Felsky. Multi-omic integration via similarity network fusion to detect molecular subtypes of ageing. *Brain Communications*, 5(2):fcad110, April 2023.

[137] Marcos Francisco Perez and Peter Sarkies. Malignancy and NF-B signalling strengthen coordination between expression of mitochondrial and nuclear-encoded oxidative phosphorylation genes. *Genome Biology*, 22:328, December 2021.

[138] amppd.org. Transcriptomics Data | AMP-PD, 2019.

[139] Yasuhiko Baba, Katerina Markopoulou, John D. Putzke, Nathaniel R. Whaley, Matthew J. Farrer, Zbigniew K. Wszolek, and Ryan J. Uitti. Phenotypic Commonalities in Familial and Sporadic Parkinson Disease. *Archives of Neurology*, 63(4):579, April 2006.

[140] Xylena Reed, Sara Bandrés-Ciga, Cornelis Blauwendraat, and Mark R. Cookson. The role of monogenic genes in idiopathic Parkinson's disease. *Neurobiology of disease*, 124:230–239, April 2019.

[141] Kimberley J. Billingsley, Ines A. Barbosa, Sara Bandrés-Ciga, John P. Quinn, Vivien J. Bubb, Charu Deshpande, Juan A. Botia, Regina H. Reynolds, David Zhang, Michael A. Simpson, Cornelis Blauwendraat, Ziv Gan-Or, J. Raphael Gibbs, Mike A. Nalls, Andrew Singleton, Mina Ryten, and Sulev Koks. Mitochondria function associated genes contribute to Parkinson's Disease risk and later age at onset. *npj Parkinson's Disease*, 5(1):1–9, May 2019. Number: 1 Publisher: Nature Publishing Group.

[142] Jin-Sung Park, Ryan L. Davis, and Carolyn M. Sue. Mitochondrial Dysfunction in Parkinson's Disease: New Mechanistic Insights and Therapeutic Perspectives. *Current Neurology and Neuroscience Reports*, 18(5):21, 2018.

[143] Theodora Ntetsika, Paraskevi-Evita Papathoma, and Ioanna Markaki. Novel targeted therapies for Parkinson's disease. *Molecular Medicine*, 27(1):17, February 2021.

[144] Anant Dadu, Vipul Satone, Rachneet Kaur, Sayed Hadi Hashemi, Hampton Leonard, Hirotaka Iwaki, Mary B. Makarious, Kimberley J. Billingsley, Sara Bandres-Ciga, Lana J. Sargent, Alastair J. Noyce, Ali Daneshmand, Cornelis Blauwendraat, Ken Marek, Sonja W. Scholz, Andrew B. Singleton, Mike A. Nalls, Roy H. Campbell, and Faraz Faghri. Identification and prediction of Parkinson's disease subtypes and progression using machine learning in two cohorts. *npj Parkinson's Disease*, 8(1):1–12, December 2022. Number: 1 Publisher: Nature Publishing Group.

[145] Per Borghammer. The -Synuclein Origin and Connectome Model (SOC Model) of Parkinson's Disease: Explaining Motor Asymmetry, Non-Motor Phenotypes, and Cognitive Decline. *Journal of Parkinson's Disease*, 11(2):455–474, January 2021. Publisher: IOS Press.

[146] Kenneth Marek, Sohini Chowdhury, Andrew Siderowf, Shirley Lasch, Christopher S. Coffey, Chelsea Caspell-Garcia, Tanya Simuni, Danna Jennings, Caroline M. Tanner, John Q. Trojanowski, Leslie M. Shaw, John

Seibyl, Norbert Schuff, Andrew Singleton, Karl Kieburtz, Arthur W. Toga, Brit Mollenhauer, Doug Galasko, Lana M. Chahine, Daniel Weintraub, Tatiana Foroud, Duygu Tosun-Turgut, Kathleen Poston, Vanessa Arnedo, Mark Frasier, Todd Sherer, and the Parkinson's Progression Markers Initiative. The Parkinson's progression markers initiative (PPMI) – establishing a PD biomarker cohort. *Annals of Clinical and Translational Neurology*, 5(12):1460–1477, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/acn3.644.

[147] Liana S. Rosenthal, Daniel Drake, Roy N. Alcalay, Debra Babcock, F. DuBois Bowman, Alice Chen-Plotkin, Ted M. Dawson, Richard B. Dewey, Dwight German, Xuemei Huang, Barry Landin, Matthew McAuliffe, Vladislav A. Petyuk, Clemens R. Scherzer, Coryse St Hillaire-Clarke, Beth-Anne Sieber, Margaret Sutherland, Chi Tarn, Andrew West, David Vaillancourt, Jing Zhang, and Katrina Gwinn. The NINDS Parkinson's Disease Biomarkers Program. *Movement disorders : official journal of the Movement Disorder Society*, 31(6):915–923, June 2016.

[148] Y. Liao, G. K. Smyth, and W. Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, April 2014.

[149] Yan Guo, Yulin Dai, Hui Yu, Shilin Zhao, David C. Samuels, and Yu Shyr. Improvements and impacts of GRCh38 human reference on high throughput sequencing data analysis. *Genomics*, 109(2):83–90, March 2017.

[150] Laura D. Osellame, Ahad A. Rahim, Iain P. Hargreaves, Matthew E. Gegg, Angela Richard-Londt, Sebastian Brandner, Simon N. Waddington, Anthony H. V. Schapira, and Michael R. Duchen. Mitochondria and Quality Control Defects in a Mouse Model of Gaucher Disease—Links to Parkinson's Disease. *Cell Metabolism*, 17(6):941–953, June 2013. Publisher: Elsevier.

[151] Mattia Vicario, Domenico Cieri, Francesca Vallese, Cristina Catoni, Lucia

Barazzuol, Paola Berto, Alessandro Grinzato, Laura Barbieri, Marisa Brini, and Tito Calì. A split-GFP tool reveals differences in the sub-mitochondrial distribution of wt and mutant alpha-synuclein. *Cell Death & Disease*, 10(11):1–16, November 2019. Number: 11 Publisher: Nature Publishing Group.

[152] David C. Schöndorf, Dina Ivanyuk, Pascale Baden, Alvaro Sanchez-Martinez, Silvia De Cicco, Cong Yu, Ivana Giunta, Lukas K. Schwarz, Gabriele Di Napoli, Vasiliki Panagiotakopoulou, Sigrun Nestel, Marcus Keatinge, Jan Pruszak, Oliver Bandmann, Bernd Heimrich, Thomas Gasser, Alexander J. Whitworth, and Michela Deleidi. The NAD+ Precursor Nicotinamide Riboside Rescues Mitochondrial Defects and Neuronal Loss in iPSC and Fly Models of Parkinson's Disease. *Cell Reports*, 23(10):2976–2988, June 2018. Publisher: Elsevier.

[153] Sylvie Delcambre, Jenny Ghelfi, Nassima Ouzren, Léa Grandmougin, Catherine Delbrouck, Philip Seibler, Kobi Wasner, Jan O. Aasly, Christine Klein, Joanne Trinh, Sandro L. Pereira, and Anne Grünewald. Mitochondrial Mechanisms of LRRK2 G2019S Penetrance. *Frontiers in Neurology*, 11, 2020.

[154] Fieke Wauters, Tom Cornelissen, Dorien Imberechts, Shaun Martin, Brianada Koentjoro, Carolyn Sue, Peter Vangheluwe, and Wim Vandenberghe. LRRK2 mutations impair depolarization-induced mitophagy through inhibition of mitochondrial accumulation of RAB10. *Autophagy*, 16(2):203–222, February 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15548627.2019.1603548.

[155] Heather Mortiboys, Ruby Macdonald, Thomas Payne, Matilde Sassani, Thomas Jenkins, and Oliver Bandmann. Translational approaches to restoring mitochondrial function in Parkinson's disease. *FEBS Letters*, 592(5):776–792, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1873-3468.12920.

[156] Tadashi Umehara, Hisayoshi Oka, Atsuo Nakahara, Hiromasa Matsuno, and Hidetomo Murakami. Differential leukocyte count is associated with clinical phenotype in Parkinson's disease. *Journal of the Neurological Sciences*, 409, February 2020. Publisher: Elsevier.

[157] Konstantinos Palikaras and Nektarios Tavernarakis. Mitochondrial homeostasis: The interplay between mitophagy and mitochondrial biogenesis. *Experimental Gerontology*, 56:182–188, August 2014.

[158] D. Grahame Hardie. AMP-activated/SNF1 protein kinases: conserved guardians of cellular energy. *Nature Reviews Molecular Cell Biology*, 8(10):774–785, October 2007. Number: 10 Publisher: Nature Publishing Group.

[159] Tommaso Schirinzi, Illari Salvatori, Henri Zenuni, Piergiorgio Grillo, Cristiana Valle, Giuseppina Martella, Nicola Biagio Mercuri, and Alberto Ferri. Pattern of Mitochondrial Respiration in Peripheral Blood Cells of Patients with Parkinson's Disease. *International Journal of Molecular Sciences*, 23(18):10863, January 2022. Number: 18 Publisher: Multidisciplinary Digital Publishing Institute.

[160] Bin Zheng, Zhixiang Liao, Joseph J. Locascio, Kristen A. Lesniak, Sarah S. Roderick, Marla L. Watt, Aron C. Eklund, Yanli Zhang-James, Peter D. Kim, Michael A. Hauser, Edna Grünblatt, Linda B. Moran, Silvia A. Mandel, Peter Riederer, Renee M. Miller, Howard J. Federoff, Ullrich Wüllner, Spyridon Papapetropoulos, Moussa B. Youdim, Ippolita Cantuti-Castelvetri, Anne B. Young, Jeffery M. Vance, Richard L. Davis, John C. Hedreen, Charles H. Adler, Thomas G. Beach, Manuel B. Graeber, Frank A. Middleton, Jean-Christophe Rochet, and Clemens R. Scherzer. PGC-1, A Potential Therapeutic Target for Early Intervention in Parkinson's Disease. *Science translational medicine*, 2(52):52ra73, October 2010.

[161] Angela Pyle, Haidyan Anugrha, Marzena Kurzawa-Akanbi, Alison Yarnall,

David Burn, and Gavin Hudson. Reduced mitochondrial DNA copy number is a biomarker of Parkinson's disease. *Neurobiology of Aging*, 38:216.e7–216.e10, February 2016.

[162] W. Lee, J. Johnson, D. J. Gough, J. Donoghue, G. L. M. Cagnone, V. Vaghjiani, K. A. Brown, T. G. Johns, and J. C. St. John. Mitochondrial DNA copy number is regulated by DNA methylation and demethylation of POLGA in stem and cancer cells and their differentiated progeny. *Cell Death & Disease*, 6(2):e1664–e1664, February 2015. Number: 2 Publisher: Nature Publishing Group.

[163] Yamilee Hurtado-Roca, Marta Ledesma, Monica Gonzalez-Lazaro, Raquel Moreno-Loshuertos, Patricio Fernandez-Silva, Jose Antonio Enriquez, and Martin Laclaustra. Adjusting MtDNA Quantification in Whole Blood for Peripheral Blood Platelet and Leukocyte Counts. *PloS One*, 11(10):e0163770, 2016.

[164] Judita Knez, Ellen Winckelmans, Michelle Plusquin, Lutgarde Thijs, Nicholas Cauwenberghs, Yumei Gu, Jan A. Staessen, Tim S. Nawrot, and Tatiana Kuznetsova. Correlates of Peripheral Blood Mitochondrial DNA Content in a General Population. *American Journal of Epidemiology*, 183(2):138–146, January 2016.

[165] Liron Ganel, Lei Chen, Ryan Christ, Jagadish Vangipurapu, Erica Young, Indraniel Das, Krishna Kanchi, David Larson, Allison Regier, Haley Abel, Chul Joo Kang, Alexandra Scott, Aki Havulinna, Charleston W. K. Chiang, Susan Service, Nelson Freimer, Aarno Palotie, Samuli Ripatti, Johanna Kuusisto, Michael Boehnke, Markku Laakso, Adam Locke, Nathan O. Stitziel, and Ira M. Hall. Mitochondrial genome copy number measured by DNA sequencing in human blood is strongly associated with metabolic traits via cell-type composition differences. *Human Genomics*, 15(1):34, June 2021.

[166] Anna V. Kotrys, Dominik Cysewski, Sylwia D. Czarnomska, Zbigniew

Pietras, Lukasz S. Borowski, Andrzej Dziembowski, and Roman J. Szczesny. Quantitative proteomics revealed C6orf203/MTRES1 as a factor preventing stress-induced transcription deficiency in human mitochondria. *Nucleic Acids Research*, 47(14):7502–7517, August 2019.

[167] Sinsuk Han, Mingyang Zhang, Yu Young Jeong, David J. Margolis, and Qian Cai. The role of mitophagy in the regulation of mitochondrial energetic status in neurons. *Autophagy*, 17(12):4182–4201, 2021.

[168] Hongyu Li, Ahrom Ham, Thong Chi Ma, Sheng-Han Kuo, Ellen Kanter, Donghoon Kim, Han Seok Ko, Yi Quan, Sergio Pablo Sardi, Aiqun Li, Ottavio Arancio, Un Jung Kang, David Sulzer, and Guomei Tang. Mitochondrial dysfunction and mitophagy defect triggered by heterozygous GBA mutations. *Autophagy*, 15(1):113–130, January 2019.

[169] Atossa Shaltouki, Chung-Han Hsieh, Min Joo Kim, and Xinnan Wang. Alpha-Synuclein Delays Mitophagy and Targeting Miro Rescues Neuron Loss in Parkinson's Models. *Acta neuropathologica*, 136(4):607–620, October 2018.

[170] Max Borsche, Sandro L. Pereira, Christine Klein, and Anne Grünewald. Mitochondria and Parkinson's Disease: Clinical, Molecular, and Translational Aspects. *Journal of Parkinson's Disease*, 11(1):45–60, 2021.

[171] Shankar J. Chinta, Jyothi K. Mallajosyula, Anand Rane, and Julie K. Andersen. Mitochondrial -synuclein accumulation impairs complex I function in dopaminergic neurons and results in increased mitophagy in vivo. *Neuroscience Letters*, 486(3):235–239, December 2010.

[172] Diego Grassi, Shannon Howard, Minghai Zhou, Natalia Diaz-Perez, Nicolai T. Urban, Debbie Guerrero-Given, Naomi Kamasawa, Laura A. Volpicelli-Daley, Philip LoGrasso, and Corinne Ida Lasmézas. Identification of a highly neurotoxic -synuclein species inducing mitochondrial damage and mitophagy in Parkinson's disease. *Proceedings of the National Academy*

of Sciences of the United States of America, 115(11):E2634–E2643, March 2018.

[173] Francois Singh, Alan R Prescott, Philippa Rosewell, Graeme Ball, Alastair D Reith, and Ian G Ganley. Pharmacological rescue of impaired mitophagy in Parkinson's disease-related LRRK2 G2019S knock-in mice. *eLife*, 10:e67604, August 2021.

[174] Laura J. Smith, Chiao-Yin Lee, Elisa Menozzi, and Anthony H. V. Schapira. Genetic variations in GBA1 and LRRK2 genes: Biochemical and clinical consequences in Parkinson disease. *Frontiers in Neurology*, 13:971252, August 2022.

[175] Darren M. O'Hara, Grishma Pawar, Suneil K. Kalia, and Lorraine V. Kalia. LRRK2 and -Synuclein: Distinct or Synergistic Players in Parkinson's Disease? *Frontiers in Neuroscience*, 14, 2020.

[176] Dan Bar-Yaacov, Idan Frumkin, Yuka Yashiro, Takeshi Chujo, Yuma Ishigami, Yonatan Chemla, Amit Blumberg, Orr Schlesinger, Philipp Bieri, Basil Greber, Nenad Ban, Raz Zarivach, Lital Alfonta, Yitzhak Pilpel, Tsutomu Suzuki, and Dan Mishmar. Mitochondrial 16S rRNA Is Methylated by tRNA Methyltransferase TRMT61B in All Vertebrates. *PLOS Biology*, 14(9):e1002557, September 2016. Publisher: Public Library of Science.

[177] Christina E. Toomey, Wendy E. Heywood, James R. Evans, Joanne Lachica, Sarah N. Pressey, Sandrine C. Foti, Mesfer Al Shahrani, Karishma D'Sa, Iain P. Hargreaves, Simon Heales, Michael Orford, Claire Troakes, Johannes Attems, Ellen Gelpi, Miklos Palkovits, Tammaryn Lashley, Steve M. Gentleman, Tamas Revesz, Kevin Mills, and Sonia Gandhi. Mitochondrial dysfunction is a key pathological driver of early stage Parkinson's. *Acta Neuropathologica Communications*, 10(1):134, September 2022.

[178] G. Arena, K. Sharma, G. Agyeah, R. Krüger, A. Grünewald, and J. C. Fitzgerald. Neurodegeneration and Neuroinflammation in Parkinson's Dis-

ease: a Self-Sustained Loop. *Current Neurology and Neuroscience Reports*, 22(8):427–440, August 2022.

[179] Cornelis Blauwendraat, Karl Heilbron, Costanza L. Vallerga, Sara Bandres-Ciga, Rainer von Coelln, Lasse Pihlstrøm, Javier Simón-Sánchez, Claudia Schulte, Manu Sharma, Lynne Krohn, Ari Siitonen, Hirotaka Iwaki, Hampton Leonard, Alastair J. Noyce, Manuela Tan, J. Raphael Gibbs, Dena G. Hernandez, Sonja W. Scholz, Joseph Jankovic, Lisa M. Shulman, Suzanne Lesage, Jean-Christophe Corvol, Alexis Brice, Jacobus J. van Hilten, Johan Marinus, The 23andMe Research Team, Johanna Eerola-Rautio, Pentti Tienari, Kari Majamaa, Mathias Toft, Donald G. Grosset, Thomas Gasser, Peter Heutink, Joshua M. Shulman, Nicolas Wood, John Hardy, Huw R. Morris, David A. Hinds, Jacob Gratten, Peter M. Visscher, Ziv Gan-Or, Mike A. Nalls, Andrew B. Singleton, and for the International Parkinson's Disease Genomics Consortium (IPDGC). Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and -synuclein mechanisms. *Movement Disorders*, 34(6):866–875, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.27659.

[180] Manuela MX Tan, Michael A Lawton, Miriam I Pollard, Emmeline Brown, Samir Bekadar, Edwin Jabbari, Regina H Reynolds, Hirotaka Iwaki, Cornelis Blauwendraat, Sofia Kanavou, Leon Hubbard, Naveed Malek, Katherine A Grosset, Nin Bajaj, Roger A Barker, David J Burn, Catherine Bresner, Thomas Foltynie, Nicholas W Wood, Caroline H Williams-Gray, Ole A Andreassen, Mathias Toft, Alexis Elbaz, Fanny Artaud, Alexis Brice, Jean-Christophe Corvol, Jan Aasly, Matthew J Farrer, Michael A Nalls, Andrew B Singleton, Nigel M Williams, Yoav Ben-Shlomo, John Hardy, Michele TM Hu, Donald G Grosset, Maryam Shoai, Lasse Pihlstrøm, and Huw R Morris. Genome-wide determinants of mortality and clinical progression in Parkinson's disease. preprint, Genetic and Genomic Medicine, July 2022.

[181] Manuela M.X. Tan, Michael A. Lawton, Edwin Jabbari, Regina H. Reynolds,

Hirotaka Iwaki, Cornelis Blauwendraat, Sofia Kanavou, Miriam I. Pollard, Leon Hubbard, Naveed Malek, Katherine A. Grosset, Sarah L. Marrinan, Nin Bajaj, Roger A. Barker, David J. Burn, Catherine Bresner, Thomas Foltynie, Nicholas W. Wood, Caroline H. Williams-Gray, John Hardy, Michael A. Nalls, Andrew B. Singleton, Nigel M. Williams, Yoav Ben-Shlomo, Michele T.M. Hu, Donald G. Grosset, Maryam Shoai, and Huw R. Morris. Genome-Wide Association Studies of Cognitive and Motor Progression in Parkinson's Disease. *Movement Disorders*, 36(2):424–433, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.28342.

[182] Brian M. Schilder, Elisa Navarro, and Towfique Raj. Multi-omic insights into Parkinson's Disease: From genetic associations to functional mechanisms. *Neurobiology of Disease*, 163:105580, February 2022.

[183] A. Williams. MPTP parkinsonism. *Br Med J (Clin Res Ed)*, 289(6456):1401–1402, November 1984. Publisher: British Medical Journal Publishing Group Section: Research Article.

[184] Elisa Navarro, Evan Udine, Katia de Paiva Lopes, Madison Parks, Giulietta Riboldi, Brian M. Schilder, Jack Humphrey, Gijsje J. L. Snijders, Ricardo A. Vialle, Maojuan Zhuang, Tamjeed Sikder, Charalambos Argyrou, Amanda Allan, Michael J. Chao, Kurt Farrell, Brooklyn Henderson, Sarah Simon, Deborah Raymond, Sonya Elango, Roberto A. Ortega, Vicki Shanker, Matthew Swan, Carolyn W. Zhu, Ritesh Ramdhani, Ruth H. Walker, Winona Tse, Mary Sano, Ana C. Pereira, Tim Ahfeldt, Alison M. Goate, Susan Bressman, John F. Crary, Lotje de Witte, Steven Frucht, Rachel Saunders-Pullman, and Towfique Raj. Dysregulation of mitochondrial and proteolysosomal genes in Parkinson's disease myeloid cells. *Nature Aging*, 1(9):850–863, September 2021. Number: 9 Publisher: Nature Publishing Group.

[185] Sarah J. Annesley, Sui T. Lay, Shawn W. De Piazza, Oana Sanislav, Eleanor Hammersley, Claire Y. Allan, Lisa M. Francione, Minh Q. Bui, Zhi-Ping

Chen, Kevin R. W. Ngoei, Flora Tassone, Bruce E. Kemp, Elsdon Storey, Andrew Evans, Danuta Z. Loesch, and Paul R. Fisher. Immortalized Parkinson's disease lymphocytes have enhanced mitochondrial respiratory activity. *Disease Models & Mechanisms*, 9(11):1295–1305, November 2016.

[186] Shi Yao, Xi Zhang, Shu-Cheng Zou, Yong Zhu, Bo Li, Wei-Ping Kuang, Yan Guo, Xiao-Song Li, Liang Li, and Xiao-Ye Wang. A transcriptome-wide association study identifies susceptibility genes for Parkinson's disease. *npj Parkinson's Disease*, 7(1):1–8, September 2021. Number: 1 Publisher: Nature Publishing Group.

[187] Jochen Ohnmacht, Patrick May, Lasse Sinkkonen, and Rejko Krüger. Missing heritability in Parkinson's disease: the emerging role of non-coding genetic variation. *Journal of Neural Transmission*, 127(5):729–748, May 2020.

[188] Xuanyao Liu, Yang I. Li, and Jonathan K. Pritchard. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*, 177(4):1022–1034.e6, May 2019.

[189] Mark Helm. Post-transcriptional nucleotide modification and alternative folding of RNA. *Nucleic Acids Research*, 34(2):721–733, January 2006.

[190] Hirotaka Iwaki, Hampton L. Leonard, Mary B. Makarious, Matt Bookman, Barry Landin, David Vismer, Bradford Casey, J. Raphael Gibbs, Dena G. Hernandez, Cornelis Blauwendraat, Daniel Vitale, Yeajin Song, Dinesh Kumar, Clifton L. Dalgard, Mahdiar Sadeghi, Xianjun Dong, Leonie Misquitta, Sonja W. Scholz, Clemens R. Scherzer, Mike A. Nalls, Shameek Biswas, and Andrew B. Singleton. Accelerating Medicines Partnership: Parkinson's Disease. Genetic Resource. *Movement Disorders*, 36(8):1795–1804, August 2021.

[191] Allison A. Regier, Yossi Farjoun, David E. Larson, Olga Krasheninina, Hyun Min Kang, Daniel P. Howrigan, Bo-Juen Chen, Manisha Kher, Eric Banks, Darren C. Ames, Adam C. English, Heng Li, Jinchuan Xing, Yeting

Zhang, Tara Matise, Goncalo R. Abecasis, Will Salerno, Michael C. Zody, Benjamin M. Neale, and Ira M. Hall. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nature Communications*, 9(1):4038, October 2018.

[192] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A. Albers, Eric Banks, Mark A. DePristo, Robert E. Handsaker, Gerton Lunter, Gabor T. Marth, Stephen T. Sherry, Gilean McVean, Richard Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, August 2011. Publisher: Oxford Academic.

[193] Florian Privé, Hugues Aschard, Andrey Ziyatdinov, and Michael G B Blum. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics*, 34(16):2781–2787, August 2018.

[194] A. A. Shabalin. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, May 2012.

[195] Regina Hertfelder Reynolds. RHReynolds/colochelpR: v0.99.1, June 2021.

[196] Yves Rosseel. **lavaan** : An *R* Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 2012.

[197] Ruchi Lohia, Nathan Fox, and Jesse Gillis. A global high-density chromatin interaction network reveals functional long-range and trans-chromosomal relationships. *Genome Biology*, 23(1):238, November 2022.

[198] David N Cooper. Functional intronic polymorphisms: Buried treasure awaiting discovery within our genes. *Human Genomics*, 4(5):284–288, June 2010.

[199] Maya Ghoussaini, Edward Mountjoy, Miguel Carmona, Gareth Peat, Ellen M Schmidt, Andrew Hercules, Luca Fumis, Alfredo Miranda, Denise Carvalho-Silva, Annalisa Buniello, Tony Burdett, James Hayhurst, Jarrod Baker, Javier Ferrer, Asier Gonzalez-Uriarte, Simon Jupp, Mohd Anisul Karim, Gautier

Koscielny, Sandra Machlitt-Northen, Cinzia Malangone, Zoe May Pendlington, Paola Roncaglia, Daniel Suveges, Daniel Wright, Olga Vrousgou, Eliseo Papa, Helen Parkinson, Jacqueline A L MacArthur, John A Todd, Jeffrey C Barrett, Jeremy Schwartzentruber, David G Hulcoop, David Ochoa, Ellen M McDonagh, and Ian Dunham. Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics. *Nucleic Acids Research*, 49(D1):D1311–D1320, January 2021.

[200] Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, Harm Brugge, Roy Oelen, Dylan H. de Vries, Monique G. P. van der Wijst, Silva Kasela, Natalia Pervjakova, Isabel Alves, Marie-Julie Favé, Mawussé Agbessi, Mark W. Christiansen, Rick Jansen, Ilkka Seppälä, Lin Tong, Alexander Teumer, Katharina Schramm, Gibran Hemani, Joost Verlouw, Hanieh Yaghootkar, Reyhan Sönmez Flitman, Andrew Brown, Viktorija Kukushkina, Anette Kalnapenkis, Sina Rüeger, Eleonora Porcu, Jaanika Kronberg, Johannes Kettunen, Bernett Lee, Futao Zhang, Ting Qi, Jose Alquicira Hernandez, Wibowo Arindrarto, Frank Beutner, BIOS Consortium, Peter A. C. 't Hoen, Joyce van Meurs, Jenny van Dongen, Maarten van Iterson, Morris A. Swertz, i2QTL Consortium, Marc Jan Bonder, Julia Dmitrieva, Mahmoud Elansary, Benjamin P. Fairfax, Michel Georges, Bastiaan T. Heijmans, Alex W. Hewitt, Mika Kähönen, Yungil Kim, Julian C. Knight, Peter Kovacs, Knut Krohn, Shuang Li, Markus Loeffler, Urko M. Marigorta, Hailang Mei, Yukihide Momozawa, Martina Müller-Nurasyid, Matthias Nauck, Michel G. Nivard, Brenda W. J. H. Penninx, Jonathan K. Pritchard, Olli T. Raitakari, Olaf Rotzschke, Eline P. Slagboom, Coen D. A. Stehouwer, Michael Stumvoll, Patrick Sullivan, Peter A. C. 't Hoen, Joachim Thiery, Anke Tönjes, Jenny van Dongen, Maarten van Iterson, Jan H. Veldink, Uwe Völker, Robert Warmerdam, Cisca Wijmenga, Morris Swertz, Anand Andiappan, Grant W. Montgomery, Samuli Ripatti, Markus Perola, Zoltan Kutalik, Emmanouil Dermitzakis, Sven Bergmann,

Timothy Frayling, Joyce van Meurs, Holger Prokisch, Habibul Ahsan, Brandon L. Pierce, Terho Lehtimäki, Dorret I. Boomsma, Bruce M. Psaty, Sina A. Gharib, Philip Awadalla, Lili Milani, Willem H. Ouwehand, Kate Downes, Oliver Stegle, Alexis Battle, Peter M. Visscher, Jian Yang, Markus Scholz, Joseph Powell, Greg Gibson, Tõnu Esko, and Lude Franke. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature Genetics*, 53(9):1300–1310, September 2021.

[201] Jingting Yu, Ming Hu, and Chun Li. Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes. *BMC Genetics*, 20(1):43, April 2019.

[202] Biola M. Javierre, Oliver S. Burren, Steven P. Wilder, Roman Kreuzhuber, Steven M. Hill, Sven Sewitz, Jonathan Cairns, Steven W. Wingett, Csilla Várnai, Michiel J. Thiecke, Frances Burden, Samantha Farrow, Antony J. Cutler, Karola Rehnström, Kate Downes, Luigi Grassi, Myrto Kostadima, Paula Freire-Pritchett, Fan Wang, BLUEPRINT Consortium, Hendrik G. Stunnenberg, John A. Todd, Daniel R. Zerbino, Oliver Stegle, Willem H. Ouwehand, Mattia Frontini, Chris Wallace, Mikhail Spivakov, and Peter Fraser. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, 167(5):1369–1384.e19, November 2016.

[203] Kishan Kumar Nyati and Tadamitsu Kishimoto. Recent Advances in the Role of Arid5a in Immune Diseases and Cancer. *Frontiers in Immunology*, 12, 2022.

[204] Michael Chong, Pedrum Mohammadi-Shemirani, Nicolas Perrot, Walter Nelson, Robert Morton, Sukrit Narula, Ricky Lali, Irfan Khan, Mohammad Khan, Conor Judge, Tafadzwa Machipisa, Nathan Cawte, Martin O'Donnell, Marie Pigeyre, Loubna Akhabir, and Guillaume Paré. GWAS and ExWAS

of blood mitochondrial DNA copy number identifies 71 loci and highlights a potential causal role in dementia. *eLife*, 11:e70382, January 2022.

[205] Elizabeth M. Kline, Laura M. Butkovich, Joshua M. Bradner, Jianjun Chang, Harris Gelbard, Val Goodfellow, W. Michael Caudle, and Malú G. Tansey. The second generation mixed-lineage kinase-3 (MLK3) inhibitor CLFB-1134 protects against neurotoxin-induced nigral dopaminergic neuron loss. *Experimental neurology*, 318:157–164, August 2019.

[206] Shashi Kant, Wojciech Swat, Sheng Zhang, Zhong-Yin Zhang, Benjamin G. Neel, Richard A. Flavell, and Roger J. Davis. TNF-stimulated MAP kinase activation mediated by a Rho family GTPase signaling pathway. *Genes & Development*, 25(19):2069–2078, October 2011. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[207] Daniel F. Marker, Marie-Ève Tremblay, Jenna M. Puccini, Justin Barbieri, Mary A. Gantz Marker, Colin J. Loweth, E. Chris Muly, Shao-Ming Lu, Val S. Goodfellow, Stephen Dewhurst, and Harris A. Gelbard. The New Small-Molecule Mixed-Lineage Kinase 3 Inhibitor URMC-099 Is Neuroprotective and Anti-Inflammatory in Models of Human Immunodeficiency Virus-Associated Neurocognitive Disorders. *The Journal of Neuroscience*, 33(24):9998–10010, June 2013.

[208] The Parkinson Study Group PRECEPT Investigators. Mixed lineage kinase inhibitor CEP-1347 fails to delay disability in early Parkinson disease. *Neurology*, 69(15):1480–1490, October 2007. Publisher: Wolters Kluwer Health, Inc. on behalf of the American Academy of Neurology Section: Articles.

[209] Julie Lotharius, Jeppe Falsig, Johan van Beek, Sarah Payne, Ralf Dringen, Patrik Brundin, and Marcel Leist. Progressive Degeneration of Human Mesencephalic Neuron-Derived Cells Triggered by Dopamine-Dependent Oxida-

tive Stress Is Dependent on the Mixed-Lineage Kinase Pathway. *The Journal of Neuroscience*, 25(27):6329–6342, July 2005.

[210] Víctor Toribio and María Yáñez-Mó. Tetraspanins interweave EV secretion, endosomal network dynamics and cellular metabolism. *European Journal of Cell Biology*, 101(3):151229, June 2022.

[211] Henar Suárez, Zoraida Andreu, Carla Mazzeo, Víctor Toribio, Aldo Emmanuel Pérez-Rivera, Soraya López-Martín, Susana García-Silva, Begoña Hurtado, Esperanza Morato, Laura Peláez, Egoitz Astigarraga Arribas, Tarson Tolentino-Cortez, Gabriel Barreda-Gómez, Ana Isabel Marina, Héctor Peinado, and María Yáñez-Mó. CD9 inhibition reveals a functional connection of extracellular vesicle secretion with mitophagy in melanoma cells. *Journal of Extracellular Vesicles*, 10(7):e12082, May 2021.

[212] Yingji Jin, Yoshito Takeda, Yasushi Kondo, Lokesh P. Tripathi, Sujin Kang, Hikari Takeshita, Hanako Kuhara, Yohei Maeda, Masayoshi Higashiguchi, Kotaro Miyake, Osamu Morimura, Taro Koba, Yoshitomo Hayama, Shohei Koyama, Kaori Nakanishi, Takeo Iwasaki, Satoshi Tetsumoto, Kazuyuki Tsujino, Muneyoshi Kuroyama, Kota Iwahori, Haruhiko Hirata, Takayuki Takimoto, Mayumi Suzuki, Izumi Nagatomo, Ken Sugimoto, Yuta Fujii, Hiroshi Kida, Kenji Mizuguchi, Mari Ito, Takashi Kijima, Hiromi Rakugi, Eisuke Mekada, Isao Tachibana, and Atsushi Kumanogoh. Double deletion of tetraspanins CD9 and CD81 in mice leads to a syndrome resembling accelerated aging. *Scientific Reports*, 8(1):5145, March 2018. Number: 1 Publisher: Nature Publishing Group.

[213] Xuan Li, Ya Feng, Xi-Xi Wang, Daniel Truong, and Yun-Cheng Wu. The Critical Role of SIRT1 in Parkinson's Disease: Mechanism and Therapeutic Considerations. *Aging and Disease*, 11(6):1608–1622, December 2020.

[214] Lucilla Parnetti, Lorenzo Gaetani, Paolo Eusebi, Silvia Paciotti, Oskar Hansson, Omar El-Agnaf, Brit Mollenhauer, Kaj Blennow, and Paolo Calabresi.

CSF and blood biomarkers for Parkinson's disease. *The Lancet Neurology*, 18(6):573–586, June 2019.

[215] David W. Craig, Elizabeth Hutchins, Ivo Violich, Eric Alsop, J. Raphael Gibbs, Shawn Levy, Madison Robison, Nripesh Prasad, Tatiana Foroud, Karen L. Crawford, Arthur W. Toga, Timothy G. Whitsett, Seungchan Kim, Bradford Casey, Alyssa Reimer, Samantha J. Hutten, Mark Frasier, Fabian Kern, Tobias Fehlman, Andreas Keller, Mark R. Cookson, and Kendall Van Keuren-Jensen. RNA sequencing of whole blood reveals early alterations in immune cells and gene expression in Parkinson's disease. *Nature Aging*, 1(8):734–747, August 2021. Number: 8 Publisher: Nature Publishing Group.

[216] Adrienne R. Henderson, Qi Wang, Bessie Meechoovet, Ashley L. Siniard, Marcus Naymik, Matthew De Both, Matthew J. Huentelman, Richard J. Caselli, Erika Driver-Dunckley, and Travis Dunckley. DNA Methylation and Expression Profiles of Whole Blood in Parkinson's Disease. *Frontiers in Genetics*, 12:640266, 2021.

[217] Nayang Shan, Zuoheng Wang, and Lin Hou. Identification of trans-eQTLs using mediation analysis with multiple mediators. *BMC Bioinformatics*, 20(3):126, March 2019.

[218] Hongjun Fu, John Hardy, and Karen E. Duff. Selective vulnerability in neurodegenerative diseases. *Nature Neuroscience*, 21(10):1350–1358, October 2018. Number: 10 Publisher: Nature Publishing Group.

[219] T. Bradley Willingham, Peter T. Ajayi, and Brian Glancy. Subcellular Specialization of Mitochondrial Form and Function in Skeletal Muscle Cells. *Frontiers in Cell and Developmental Biology*, 9:757305, 2021.

[220] Timothy R. Koves, Robert C. Noland, Andrew L. Bates, Sarah T. Henes, Deborah M. Muoio, and Ronald N. Cortright. Subsarcolemmal and intermyofibrillar mitochondria play distinct roles in regulating skeletal muscle

fatty acid metabolism. *American Journal of Physiology. Cell Physiology*, 288(5):C1074–1082, May 2005.

[221] Timothy P O'Leary, Kaitlin E Sullivan, Lihua Wang, Jody Clements, Andrew L Lemire, and Mark S Cembrowski. Extensive and spatially variable within-cell-type heterogeneity across the basolateral amygdala. *eLife*, 9:e59003, September 2020. Publisher: eLife Sciences Publications, Ltd.

[222] Mike Fainzilber, Vivian Budnik, Rosalind A. Segal, and Michael R. Kreutz. From Synapse to Nucleus and Back Again—Communication over Distance within Neurons. *Journal of Neuroscience*, 31(45):16045–16048, November 2011. Publisher: Society for Neuroscience Section: Symposium and Mini-Symposium.

[223] Rodrigo C. Vergara, Sebastián Jaramillo-Riveri, Alejandro Luarte, Cristóbal Moënne-Loccoz, Rómulo Fuentes, Andrés Couve, and Pedro E. Maldonado. The Energy Homeostasis Principle: Neuronal Energy Regulation Drives Local Network Dynamics Generating Behavior. *Frontiers in Computational Neuroscience*, 13:49, July 2019.

[224] Binoy Paulose Nadappuram, Paolo Cadinu, Avijit Barik, Alexander J. Ainscough, Michael J. Devine, Minkyung Kang, Jorge Gonzalez-Garcia, Josef T. Kittler, Keith R. Willison, Ramon Vilar, Paolo Actis, Beata Wojciak-Stothard, Sang-Hyun Oh, Aleksandar P. Ivanov, and Joshua B. Edel. Nanoscale tweezers for single-cell biopsies. *Nature Nanotechnology*, 14(1):80–88, January 2019.

[225] Diana Iulia Savu and Nicoleta Moisoi. Mitochondria - Nucleus communication in neurodegenerative disease. Who talks first, who talks louder? *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1863(7):148588, October 2022.

[226] Vinod Udayar, Yu Chen, Ellen Sidransky, and Ravi Jagasia. Lysosomal dys-

function in neurodegeneration: emerging concepts and methods. *Trends in Neurosciences*, 45(3):184–199, March 2022. Publisher: Elsevier.

[227] Yvette C. Wong, Soojin Kim, Wesley Peng, and Dimitri Krainc. Regulation and Function of Mitochondria–Lysosome Membrane Contact Sites in Cellular Homeostasis. *Trends in Cell Biology*, 29(6):500–513, June 2019. Publisher: Elsevier.

[228] Lorena Fernández-Mosquera, Cátia V. Diogo, King Faisal Yambire, Gabriela L. Santos, Marta Luna Sánchez, Paule Bénit, Pierre Rustin, Luis Carlos Lopez, Ira Milosevic, and Nuno Raimundo. Acute and chronic mitochondrial respiratory chain deficiency differentially regulate lysosomal biogenesis. *Scientific Reports*, 7:45076, March 2017.

[229] Chun Chen, Doug M. Turnbull, and Amy K. Reeve. Mitochondrial Dysfunction in Parkinson's Disease—Cause or Consequence? *Biology*, 8(2):38, May 2019.

[230] Patricia González-Rodríguez, Enrico Zampese, Kristen A. Stout, Jaime N. Guzman, Ema Ilijic, Ben Yang, Tatiana Tkatch, Mihaela A. Stavarache, David L. Wokosin, Lin Gao, Michael G. Kaplitt, José López-Barneo, Paul T. Schumacker, and D. James Surmeier. Disruption of mitochondrial complex I induces progressive parkinsonism. *Nature*, 599(7886):650–656, November 2021.

[231] Minee L. Choi, Alexandre Chappard, Bhanu P. Singh, Catherine Maclachlan, Margarida Rodrigues, Evgeniya I. Fedotova, Alexey V. Berezhnov, Suman De, Christopher J. Peddie, Dilan Athauda, Gurvir S. Virdi, Weijia Zhang, James R. Evans, Anna I. Wernick, Zeinab Shadman Zanjani, Plamena R. Angelova, Noemi Esteras, Andrey Y. Vinokurov, Katie Morris, Kiani Jeacock, Laura Tosatto, Daniel Little, Paul Gissen, David J. Clarke, Tilo Kunath, Lucy Collinson, David Klenerman, Andrey Y. Abramov, Mathew H. Horrocks, and Sonia Gandhi. Pathological structural conversion of -synuclein at the mito-

chondria induces neuronal toxicity. *Nature Neuroscience*, 25(9):1134–1148, September 2022.