Studies of gene expression in the Parkinson's disease brain

Fiona Dick

Thesis for the degree of Philosophiae Doctor (PhD)
University of Bergen, Norway
2021

UNIVERSITY OF BERGEN

Studies of gene expression in the Parkinson's disease brain

Fiona Dick

Thesis for the degree of Philosophiae Doctor (PhD)
at the University of Bergen

Date of defense: 01.10.2021

Year: 2021

Title: Studies of gene expression in the Parkinson's disease brain

Name: Fiona Dick

Print: Skipnes Kommunikasjon / University of Bergen

Scientific environment

This study was carried out in the Neuromics group led by Prof. Dr. Charalampos Tzoulis in the Institute of Clinical Medicine at the University of Bergen, Norway and the Neuro-SysMed center of excellence for clinical research in neurological diseases. The work is supported by grants from The Research Council of Norway (288164, 240369) and Bergen Research Foundation (BFS2017REK05). Both of these were received by Charalampos Tzoulis. The work was supervised by Prof. Dr. Charalampos Tzoulis and Dr. Gonzalo S. Nido.

Für Dich, Mama.

Acknowledgements

My sincere appreciation to the patients and control subjects who participated in these studies.

I am deeply grateful for the opportunity to pursue my PhD within the Neuromics group, led by my main supervisor Prof. Dr. Charalampos (Haris) Tzoulis. The atmosphere in this scientific environment he built up, is what I enjoyed most during these three years. His work effort, ambition and curiosity are infectious and have inspired and motivated me in times when I might have lacked confidence. Haris, your knowledge and expertise are most impressive and I feel so honoured to have had you as my supervisor. Thank you also for being my friend and offering your support whenever needed.

I want to thank my second supervisor Dr. Gonzalo Nido (Gon) for enabling me the best PhD experience I could have wished for. Your scientific knowledge covering all aspects of this work and your excellent guidance were essential. Together with Haris' guidance you certainly contributed to my personal development throughout those years. Thank you for always taking time to discuss "crazy" hypotheses, for always offering your help and for sharing excitement about new ideas with me. I am also more than grateful for our friendship and cannot wait to continue working with you.

Our little Bioinformatics office would not be the same without you Lilah. Thanks also to you, for your guidance and the most valuable discussions. You were always ready to help and invest time, both work- and non-work related. I am beyond happy to have you as a friend and colleague.

I want to express my gratitude to all my co-authors for their guidance, help and collaboration. Particularly, I want to thank Christian Doelle for his contribution to the DTU paper and his valuable and constructive advice on my thesis.

Working in the Neuromics group is definitely a great experience, in particular, when no COVID restrictions apply and we can enjoy some togetherness. I would like to thank all my colleagues (Birgitte, Brage, Chris, Dagny-Ann, Gia, Gon, Gry Hilde, Hanne Linda, Irene, Janani, Kristoffer, Lilah, Martina, Melli, Nelson, Romain, Thomas). I am looking forward to upcoming real-life events together.

Despite being far away from family and friends, I never felt separated from them. Thank you to my dear friends Nina, Katy, Melli and Clara for your support and help in this

period of my life. Thank you for always offering me a bed in Berlin (and Heidenheim after too much drinking :D) and for visiting me wherever I move and for being the good friends you are. I would like to thank you Konni, for our beautiful daughter, for being such a big part of my life for a long time, for always supporting me and for making theses seemingly impossible logistic arrangements possible. Thanks also to your parents (family Szengel) for helping me in all ways possible, for their visits and for taking care of Finja when needed.

I would like to express my gratitude to my family (my mum Birgitta, my dad Wolfram and my sister Marlien), who never questioned my abilities and supported me and my decisions throughout my complete education, even if these involved me moving far away. Danke Mama, Papa, Lini und Oma Kathi. My deepest admiration goes to my grandfather Heini Dick, who, since I was little, impressed me with his intelligence and inspired me with his insatiable curiosity and infinite urge to learn and invent. Danke Opa, fuer deine konstante Unterstuetzung.

Jonathan, you came into my life when things were not necessarily easy and quite stressful and ironically you are the epitome of calmness and patience. I don't know what I would have done without you and I surely cannot wait to experience my life with you part of it. You are the most beautiful sunflower. I am so excited to meet our child in July, who has been with me throughout this last stretch of writing the thesis.

Lastly, I am most proud to have you, Finja, as my daughter. I am lacking words to describe my love and appreciation for you. I enjoy every day with you and being able to witness your learning and developing is the greatest gift I could imagine. I really hope you never stop wanting to hold my hand while we sleep and never stop telling me that I am "die ganze Liebe der Welt" :).

Abstract

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder, affecting $\sim 1.8\%$ of the population above 65 years. A combination of genetic and environmental factors contributes to the risk of PD, but the molecular mechanisms underlying its aetiology remain largely unaccounted for.

Profiling gene expression in the PD brain can identify molecular processes associated with the pathogenesis and nominate candidate therapeutic targets for further study. Most previous gene expression studies in PD focused on specific hypotheses and were restricted to selected genes of interest and only few were performed transcriptome-wide. While in part informative, the results of these studies must be interpreted with caution due to a combination of technical and biological limitations. Factors applying specifically to the study of human bulk brain tissue make it difficult to confidently and accurately determine altered pathways. 1) Bulk brain tissue is composed of multiple cell types, some of which are selectively affected in PD. Variation in cell-type composition across samples introduces noise, while disease-associated changes in the number of neurons and glia introduce systematic gene expression biases between conditions. 2) The complex architecture of neurons complicates sample dissection and can result in variable soma-to-synapses ratios across samples. This variability results in additional noise in expression data since RNA and proteins can undergo axonal transport, with some preferentially localizing to the soma or synapses. Another limitation of previous studies is that gene-level analyses provide only an incomplete perspective on the expression landscape. Regulation at the transcript- and protein-level is often overlooked.

The work of this thesis comprises three alternative approaches of gene expression analyses in the PD brain, aiming to overcome these limitations. We employed RNA-Seq and mass spectrometry in the prefrontal cortex of PD patients and healthy controls and approached these challenges by profiling expression at transcript-, gene- and protein-level. Considering the described aspects of bulk brain tissue, we adjusted for changes in cellular composition, RNA quality and guided functional interpretation with the polarized nature of neurons in mind.

Our results indicate that the frequently reported downregulation of mitochondrial function is partly driven by cellular composition. Adjusting for cell-type bias instead revealed

altered pathways related to protein degradation, further strengthening their involvement in disease pathology. Both differential gene and transcript isoform expression showed enrichment for these. Additionally, we nominated genes that exhibit differential transcript usage events, suggesting alternate regulation at the transcript-level. These candidates can be targeted in future studies to identify functional consequences. Finally, we observed discordance between transcriptome and proteome which we concluded reflects alterations in PD proteostasis. Specifically, we identified certain proteasomal subunits central to these regulatory changes, providing us with further evidence for the key role of protein degradation in PD brain.

List of publications

The papers included and discussed in this thesis are:

1. Gonzalo S. Nido, Fiona Dick, Lilah Toker, Kjell Petersen, Guido Werner Alves, Ole-Bjørn Tysnes, Inge Jonassen, Kristoffer Haugarvoll & Charalampos Tzoulis , (2020) *Common gene expression signatures in Parkinson's disease are driven by changes in cell composition* , Acta Neuropathologica Communications, **8/55(2020)**

2. Fiona Dick, Gonzalo S. Nido, Guido Werner Alves, Ole-Bjørn Tysnes, Gry Hilde Nilsen, Christian Dölle, Charalampos Tzoulis, 2020 *Differential transcript usage in the Parkinson's disease brain*, PLOS Genetics, **16(11):e1009182**

3. Fiona Dick, Guido Alves, Ole-Bjørn Tysnes, Gonzalo S. Nido, Charalampos Tzoulis, *Altered transcriptome-proteome coupling indicates aberrant proteostasis in Parkinson's disease*, Manuscript

List of abbrevations

AD	Alzheimer's disease
bp	Base pairs
DNA	Deoxyribonucleic acid
cDNA	Complementary DNA
DA	Dopaminergic
DEG	Differentially expressed gene
DGE	Differential gene expression
DPE	Differential protein expression
DTE	Differential transcript expression
DTU	Differential transcript usage
ER	endoplasmic reticulum
GWAS	Genome wide association study
HPLC	High pressure liquid chromatography
LC	Liquid chromatography
LCM	Laser capture microdissection
LP	Lewy pathology
MGP	Marker gene profile
MRC	Mitochondrial respiratory chain
mRNA	Messenger RNA
MS	Mass spectroscopy
PC	Principal component
PCC	Posterior cingulate cortex
PD	Parkinson's disease
PFC	Prefrontal cortex
poly-A	polyadenylated
PTM	Post transcriptional modifications
RIN	RNA integrity number
RNA-Seq	RNA sequencing
rRNA	ribosomal RNA
SNc	Substantia nigra pars compacta

STR	Striatum
TMT	Tandem mass tags
UPS	Unfolded protein response
UPS	Ubiquitin-proteasome system
UTR	Untranslated region
qPCR	Quantitative polymerase chain reaction

Contents

Chapter 1

Introduction

This chapter introduces the scientific background of this work by summarizing principles of molecular biology, characterizing Parkinson's disease (PD), defining methodological aspects and reviewing the current knowledge on gene expression in the PD brain.

1.1 Gene expression

In 1909, Wilhelm Johannsen introduced the term *gene* to describe the units associated with inherited traits [92]. The human genome, consisting of base-paired nucleotides which form the double-stranded DNA (deoxyribonucleic acid) molecule, is now estimated to comprise over 3 billion base pairs (bp) of genetic code, organized into 23 pairs of chromosomes [168] in the nucleus of the cell.
In 1958, Francis Crick introduced the term central dogma of molecular biology [39] and described a "residue-by-residue transfer of sequential information" from DNA to the intermediate messenger RNA (*transcription*) and from RNA to the protein (*translation*). In 1970, he expanded on the explanation of his definition in response to critique labelling his theory as oversimplified [38]. He suggested information transfers from DNA to RNA, from RNA to protein, from DNA to DNA, from RNA to RNA, and possibly from RNA to DNA (the latter two appeared in virus-infected cells) (Figure 1.1A).
The theory of an *injective* central dogma (i.e., a function that maps each element (gene) to exactly one image (protein)), implies that the diversity of proteins is explained by the nucleic acid sequence variation [177]. This incomplete view has long since been expanded to a more complex understanding of how genes encode proteins. Proteins are no longer viewed as the only functional end product of a gene. Instead, different types of non-protein-coding RNA (Figure 1.1B), such as ribosomal or micro RNA are known to

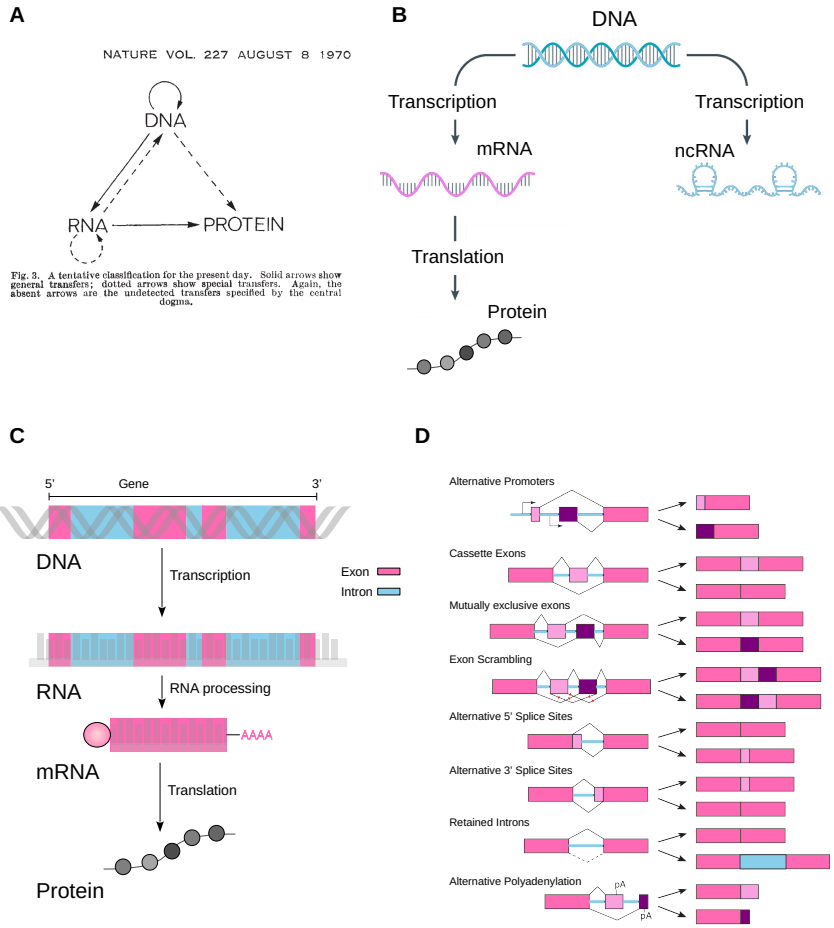


Figure 1.1: **General aspects of gene expression**
A) Extract of "Central Dogma of Molecular Biology" (reprinted from [38]). **B)** Schematic diagram of possible outcomes of transcription: mRNA and non-protein-coding RNA (adapted from [172]). **C)** Schematic presentation of a simplified protein-coding gene template, its RNA transcript and its primary transcript isoform. **D)** Examples of different types of alternative splicing. Exons are represented in pink shades, introns in light blue. Unprocessed RNA is visualized on the left side, processed RNA on the right (adapted from [29])

exhibit functionality. The fate of these non-translated RNA transcripts is diverse and often regulated (e.g., function, cellular localization, half-life) [174].

A protein-coding gene template is comprised of coding DNA segments (exons), interspersed among non-coding regions (introns). The complete sequence of exons and in-

trons, together with flanking regulatory regions (termed 5'- and 3' untranslated regions (UTRs)) are transcribed in the nucleus of the cell by RNA polymerase II (in eukaryotes) to the single-stranded pre-messenger RNA (pre-mRNA) in a process called transcription. Introns are removed from the pre-mRNA co- or post-transcriptionally by a process termed *splicing* to form the final messenger RNA (mRNA). Further transcript processing involves the addition of a 5'-cap and a 3'-poly-A tail. The processed mRNA is then transported to the ribosome, where it is translated into an amino acid sequence by a process termed *translation* (Figure 1.1C). Finally, the peptide will fold into a 3-dimensional structure to exert its biological function [37].

Each step of this process is highly regulated: i) at the genetic and epigenetic level, influencing initiation and rate of transcription ii) at the transcript level, impacting the fate of the transcribed RNA molecule and iii) at the protein level, affecting protein function, localization and half-life. At the genetic level, sequence motifs and DNA binding proteins control both the initiation and rate of transcription. Non-coding regulatory sequences localized up-and down-stream of the gene‘s promoter site interact with intermediary factors to enhance or repress the binding of transcription initiation factors and RNA synthesis by the RNA polymerase [130, 37].

At the epigenetic level, mechanisms like DNA methylation and histone modification control gene expression by regulating the binding of transcription activators and inhibitors, and the state of chromatin accessibility. For example, DNA methylation is typically associated with repression of gene expression, whereas histone acetylation commonly renders chromatin accessible, thereby leading to increased transcription [45]. The epigenome is sensitive to environmental factors, and epigenetic modulation influences disease risk and progression via alternate regulation of gene expression [132, 183, 11].

Regulation also takes place at the transcript level. One major regulatory process that targets the RNA transcript itself is *alternative* splicing. It is estimated to occur in approximately 95% of genes comprising more than one exon [91] and is highly tissue-specific [174]. A multitude of alternative splicing events has been characterized. Instead of the canonical splicing of exons, these can be combined for example by skipping exons and/or including introns (Figure 1.1D). Other RNA processing mechanisms like alternative cleavage and polyadenylation or varying 3'-untranslated regions are also common [54]. Additionally, post-transcriptional regulation (before translation) can involve the interaction between the RNA and regulators such as micro RNAs [57] or RNA binding proteins [65].

These transcript-specific mechanisms regulate the composition of the transcriptome. It comprises RNA transcripts of diverse stability and half-life, sub-cellular localization and functionality. Transcriptional regulation ultimately also affects the rate and initiation of translation of protein-coding RNA, thereby shaping the proteome [54]. Finally, alternative RNA transcript processing can result in transcript variants that are non-protein-

coding or in transcript variants that encode protein isoforms with altered functionality (Figure 1.1B).
At the protein level, another layer of regulation is achieved for example by post-translational modifications (PTMs) through covalent addition of functional groups, which regulate protein function, stability, localization and degradation. Major PTMs include phosphorylation, acetylation, and glycosylation, but also hydroxylation and C-terminal amidation (for example of peptide hormones). Another possibility of protein processing is cleavage of the protein, for example, to remove a targeting signal after arrival at the subcellular destination, or generation of a peptide hormone [169].
A balanced state of the proteome, depending on protein synthesis, protein folding, stability, and protein degradation can be summarized by the term protein homeostasis or *proteostasis* [84]. The maintenance of proteostasis is highly regulated, including the initiation and rate of translation, enzyme-assisted folding of the amino acid chain, transport of proteins to their target location, unfolding of proteins and finally the rate of their biological degradation [10]. An aberrant proteostasis can lead to the accumulation of misfolded and non-functional protein.
In general, the complex regulatory network of gene expression enables tissue and cell specificity and contributes to differentiation and development and changes with human disease and ageing [150, 151, 103, 131, 91]. Studies of gene expression, in relevant cells and tissues, allow us to identify disease-associated biological processes which can be further investigated to understand disease initiation and progression and to develop targeted therapies.

1.2 Studying gene expression

Experimental studies of gene expression focus on qualitative and/or quantitative assessment of RNA and protein. Early low-throughput gene expression profiling methods like northern blots [171] and quantitative polymerase chain reaction (qPCR) [166] lacked the possibility of parallelization [50] and were restricted to the detection and quantification of only a few genes per assay in each sample. With the development of hybridization-based microarray assays, the characterization of genome-wide expression patterns was rendered possible through hybridization of pre-defined DNA probes with synthesized fluorescently labelled complementary DNA (cDNA). Microarrays remained popular gene expression methods until they were superseded by sequencing-based methodologies like RNA-sequencing (RNA-Seq) [175]. The breakthrough of next-generation DNA sequencing technologies making genome-wide expression quantification through cDNA sequencing possible, omitting the need for pre-defined probe sets, enabled de-novo transcript

detection and, most importantly, empowered hypothesis-free analyses.
Similarly, protein studies evolved from low throughput, targeted approaches based on specific antibody binding, such as Western blot [100] and ELISA [109], to high-throughput, hypothesis-free methodologies targeting the entire proteome by mass spectroscopy (MS) and related technologies [9]. Here, I will briefly describe common experimental and computational approaches involved in high-throughput RNA-Seq analyses, a central methodology to the work presented in this thesis. In addition, I will give a brief introduction to high throughput proteomics.

1.2.1 Data generation approaches

RNA sequencing

RNA-Seq is performed using high throughput sequencing technologies like the popular Illumina sequencing by synthesis. The general processes briefly summarized here.
Following isolation, RNA undergoes fragmentation, followed by cDNA synthesis to create a library of cDNA fragments, each with adapters attached to their 5'- and 3'-ends. cDNA fragments are loosely attached to a flow cell through hybridization with tethered complementary adapters. A complementary strand of the hybridized fragment is created, starting from the complementary adapter that is solidly attached to the flow cell. The two strands are then denatured and the original cDNA fragment washed away. The newly synthesized fragment is attached to the flow cell through the adapter to which the original fragment was hybridized. Next, each DNA molecule is amplified in a process called cluster generation. The synthesized fragments hybridize with their top adapters to surrounding complementary adapters forming a bridge and allowing for amplification by creating additional complementary strands, which, after denaturation are both separated and tethered to the flow cell. This process is repeatedly performed for all fragments simultaneously, finally resulting in clone amplified fragments. After the removal of reverse strands, during sequencing-by-synthesis, from one or both ends in an additional sequence cycle (paired-end sequencing), a characteristic fluorescent light is emitted for each nucleotide that is added to the strand, resulting in clusters of fragments to light up in the same colour, until the synthesis of the short sequence (read) is completed [12]. With the clonal amplification, this sequencing methodology also termed "ensemble-based" provides *relative* quantification of the expression, which possesses the advantage of robustness against single mismatches, thereby reducing the sequencing error rate to $< 1\%$ [99].
Highly parallel sequencing of samples is achieved through multiplexing, whereby multiple samples can be sequenced in the same lane of the flow cell with the help of distinct

sample indices embedded in RNA molecule adapters.
Prior to library preparation, the collection of isolated RNA can be subject to filtering of specific species of RNA molecules, for example by enrichment via the polyadenylated (poly-A) 3'-ends of mature processed, protein-coding, mRNA, or through the removal of ribosomal RNA (rRNA). Both procedures ensure appropriate signal detection by avoiding the sequencing of highly abundant rRNA. While poly-A selection distinctly detects protein-coding mRNA, it lacks the possibility to study non-coding RNAs such as lncRNAs. Additionally, accurate transcript quantification is highly dependent on the integrity of the RNA. Due to the tail selection, partially degraded transcripts exhibit a 3'-end bias (i.e., the read coverage differs between the 5' (fewer reads) and 3' (more reads) ends) [185].

The sequencing methodology described above is often termed "short-read sequencing", since its output consists of relatively short reads ($\leq$ 200 bp) with an average library size of 20-30 million reads per sample [157]. Depending on the complexity of the transcriptome of interest and the chosen RNA selection methodology, more reads might be necessary for accurate quantification of both low and abundantly expressed transcripts and to achieve satisfying coverage of the transcriptome (average genome coverage $= \frac{\text{number of reads} \cdot \text{average read length (bp)}}{\text{genome or transcriptome length (bp)}}$ [146]).
Following alignment of raw sequence reads to a reference genome or transcriptome, gene expression is quantified by the number of reads mapped to annotated regions defined as genes. The alignment of reads from RNA-Seq experiments to a reference genome will result in split reads on exon-exon junctions, where the reference genome constitutes an intron, which is not present in the sequenced mature RNA due to splicing. Alignment and quantification constitute the main computational bottlenecks in RNA-seq processing, which has led to the active development of advanced computational methods to reduce run-time. For example, recently developed algorithms reduce the steps of single base alignment and quantification to infer transcript abundance directly through pseudo- or quasi-mapping [21, 136]. Difficulties, however, remain: ambiguously mapped reads, pose the challenge of determining their origin. Distinct alignment of reads to transcript isoforms arising from alternative splicing or other alternative transcript processing is not trivial.

Proteomics

Proteomics is the qualitative and/or quantitative study of the proteome, i.e., the collection of proteins present in a cell, tissue or organism. The proteomics field has evolved a broad arsenal of technologies and methodologies, ranging from relative and absolute protein quantification to the detection and quantification of PTMs. While methodologically

diverse and tailored to the research question at hand, proteomic analyses are generally based on the following broad principles. First, protein fraction is purified from the sample of interest, digested into peptides, commonly by trypsinization, and either fractionated according to their electrochemical properties (e.g., size and charge) or enriched for specific peptide subpopulations (e.g., specific PTMs). The fractionated or enriched peptide populations are separated, commonly using a liquid chromatography (LC)-based technique (e.g., high-pressure liquid chromatography, HPLC). The separated peptides are then ionized and assessed by mass spectrometry to assess their mass/charge ratio and determine their identity. Identified peptides are quantified in relative or absolute terms, depending on the design of the experiment. Finally, bioinformatic approaches are employed to assemble, quality control and analyze the data [2]. Parallel MS-analysis of multiple samples is possible via multiplexing methodologies, such as "tandem mass tags" (TMT) [164]. While powerful and constantly evolving, currently mainstream proteomic methodologies with the inclusion of multiple batches challenged by a high abundance of missing values, batch effects, and high false-positive rates introduced through channel leakage, influencing both sensitivity and specificity of expression quantification [22].

1.2.2 Data analysis approaches

Differential gene expression and pathway enrichment analysis

In differential gene expression (DGE) analyses, transcript-level expression counts are aggregated to the gene-level and then compared between conditions. Thus, DGE analysis investigates changes in abundance of the summarized transcriptional output of a gene. Differentially expressed genes (DEGs) can then be ranked by the magnitude of change in expression or by its level of significance. This ranking can be used to test for the enrichment/overrepresentation in specific functions.

Studying transcript isoforms: differential transcript usage

One way to study differences in the expression of transcript isoforms is differential transcript usage (DTU). In contrast to DGE, for DTU analysis, estimated transcript counts are not aggregated at the gene level. This enables the consideration of potentially functionally distinct transcript isoforms which originate from the same gene template. A direct comparison of expression of transcript isoforms between two conditions is referred to as differential transcript expression (DTE). For DTU, however, transcript isoform abundance is investigated in relation to the complete transcriptional output of a gene

(Figure 1.2). Thus, DTU analyses estimate "transcript usage" and detect changes in the relative contribution of a transcript to the overall expression of the gene. Transcript usage corresponds to transcript-level expression counts of a transcript i normalized by the sum of counts of all transcripts of a gene j:

$$TU_{i,j} = \frac{t_i}{\sum_{k=1}^{n_j} t_k} \ , \tag{1.1}$$

where n_j equals the number of transcripts of gene j and t_i is the expression count of transcript i. Hence, *differential* transcript usage describes a change in proportions between two conditions.

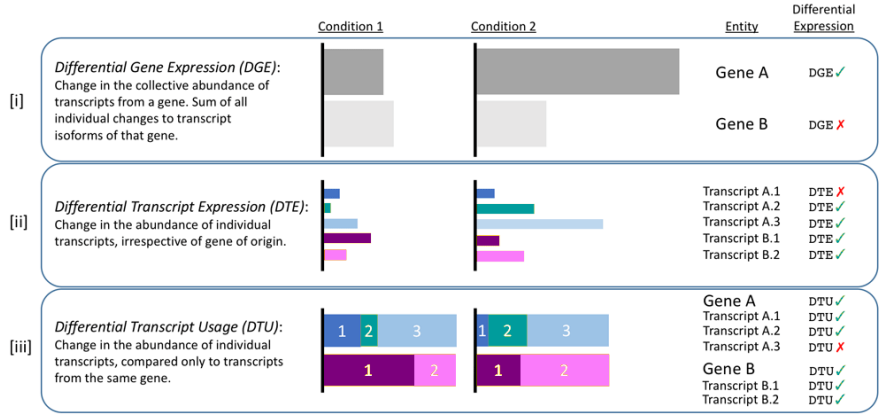


Figure 1.2: **Definition of DGE, DTE and DTU** Schematic illustration of types of differential expression (reprinted and adapted from [61])

Differential protein expression

Depending on the employed proteomics methodology, acquired peptide intensities are aggregated to protein level abundances. Like in DGE analysis, differential protein expression (DPE) analysis identifies differences in protein levels between two conditions, and subsequent functional enrichment analysis can be performed on the preferred statistic of differentially expressed proteins (e.g., fold-change or significance)

1.3 Parkinson's disease

1.3.1 Epidemiology and clinical features

Parkinson's disease (PD) is a major cause of death and disability and has a devastating global socioeconomic impact. It affects 1-2% of the population above the age of 65 years and its prevalence increases as the population ages [44, 66]. In Europe alone, PD affects ∼1.2 million people and has an estimated cost of €14 billion/year [70]. Despite more than two centuries of research, the molecular mechanisms underlying PD remain largely unknown and there are no disease-modifying therapies able to prevent or delay disease initiation and progression. Thus, the need to understand and treat PD has never been more urgent. While references to parkinsonian symptoms can be found since ancient records, including an essay written by Galen in 169 CE on forms of tremor, the first systematic medical description of the PD syndrome was published by James Parkinson in 1817 in his work entitled "An Essay on the Shaking Palsy" [134]. The incidence and prevalence of PD are strongly dependent on age. PD has an estimated prevalence of 0.3% of the entire population in developed countries, but this number varies greatly between age-groups [129]. While rare below the age of 50 years, PD is estimated to affect ∼1.8% of the population above the age of 65 years and more than 4% of those over 85 years [44]. Reported incidence rates vary between 8-18 per 100,000 population per year [43] and this variation reflects both methodological and population-specific age composition variability. The sexes are asymmetrically affected with a male to female ratio of ∼1.5.

The clinical constellation of PD comprises a combination of motor and non-motor features. Motor symptoms include resting tremor, bradykinesia, rigidity and postural instability. Typical non-motor symptoms include neuropsychiatric dysfunction, olfactory loss, autonomic dysregulation, gastrointestinal dysmotility, sleep disorders, cognitive impairment and dementia[94]. Current treatments for PD are purely symptomatic and provide partial and transient relief for some of the motor symptoms. Since disease-modifying neuroprotective therapies are lacking, the neurodegenerative process of PD develops inexorably, leading to progressive disability and premature death.

1.3.2 Aetiology

PD can be divided according to aetiology into monogenic and idiopathic forms. The term PD will henceforth refer to idiopathic PD unless otherwise specified. Monogenic forms of

PD account for generally less than 5% of all cases [14]. Multiple genes have been linked to monogenic PD. Mutations in *SNCA*, *LRRK2* and *VPS35* cause autosomal dominant disease, whereas mutations in *PRKN*, *PARK7*, *PINK1*, *ATP13A2*, *FBXO7*, *PLA2G6*, *DNAJC6*, *SYNJ1* and *VPS13C* are causes of autosomal recessive forms of PD and/or parkinsonism.
Idiopathic forms of PD cannot be linked to a single causal gene and are believed to be caused by a yet unresolved interplay between genetic predisposition and environmental influences. The genetic background contributes to the risk of idiopathic PD, as evident from twin studies consistently detecting a higher risk for monozygotic than for dizygotic twins [180]. Furthermore, multiple studies have shown familial aggregation of PD, with the relative risk for PD patients to have a first degree relative with PD generally estimated between ∼2-3 % [63, 163]. The estimated overall heritability of idiopathic PD varies substantially between studies and is thought to be ∼30-40 % [73, 180]. Genome wide association studies (GWAS) have identified multiple genetic loci associated with idiopathic PD, including variation linked to *SNCA*, *LRRK2*, *GBA*, *MAPT*, and *HLA*. The latest meta-analysis [127] identified a total of 90 variants in 78 loci, which collectively explained 16-36% of the heritable risk of PD. The cause of the remaining heritable risk of PD remains unknown and is commonly referred to as the "missing heritability".
As heritability estimates and GWAS suggest, most of the risk of PD is not inherited and is, therefore, assumed to be related to external, environmental factors. In spite of extensive epidemiological studies, only a few environmental factors have been definitively linked to the disease and appear to have overall weak effects. Tobacco smoking, consumption of caffeinated beverages (including coffee and tea), and high serum urate levels have been associated with a reduced risk of PD [80]. Conversely, consumption of dairy products, exposure to pesticides (in particular paraquat and rotenone) and certain heavy metals, rural living, and traumatic brain injury have been repeatedly associated with an increased risk of PD [7]. For many of these factors, findings are conflicting, however, and whether they have a causal role in PD remains to be determined.

1.3.3 Pathology

The pathological hallmark of PD is the loss of the dopaminergic neurons of the *substantia nigra pars compacta* (SNc) in the presence of intraneuronal inclusions, composed primarily of insoluble aggregates of α-synuclein, which are collectively termed Lewy pathology. In addition, PD is characterized by progressive neuronal loss across multiple regions in the nervous system, including the olfactory bulb, amygdala, hippocampus, basal cholinergic nuclei, hypothalamus, multiple brainstem nuclei and the autonomic and enteric nervous systems [48, 47]. Other typical features include reactive changes in

the form of astrocytic gliosis and microglial activation, commonly accompanying neuronal loss, and the variable presence of other pathological protein aggregates, such as tau-containing neurofibrillary tangles and beta-amyloid plaques [48, 47]. Yet another typical pathological feature of PD is qualitative mitochondrial respiratory chain (MRC) changes, particularly in the form of neuronal respiratory chain complex I (termed henceforth "complex I") deficiency, which was first reported in the SNc [143, 74] and later found to be widespread throughout the PD brain [58].
While the neuropathological features of PD have been extensively described, their link to neuronal dysfunction and clinical impairment remains poorly understood. Thus, it is not known which of these pathological changes contribute to disease and should be targeted by therapies

1.3.4 Molecular pathophysiology

While the cascade of molecular and cellular events underlying initiation and progression of PD remain largely unknown, several biological processes have been strongly associated with the disease. Some of the most widely studied include α-synuclein aggregation, impaired proteostasis, and mitochondrial dysfunction.

Abnormal protein aggregation

α-synuclein aggregation and Lewy pathology (LP) formation are the pathological hallmarks of PD. Normally, α-synuclein exists as soluble monomers and multimers localized predominantly in presynaptic terminals [83]. In PD, due to poorly understood processes, α-synuclein undergoes misfolding and forms insoluble fibrils. These aggregate into toxic oligomers and gradually take the form of histologically recognizable aggregates termed Lewy bodies and neurites [83]. It is hypothesized that oligomeric and/or aggregated forms of α-synuclein are toxic to neurons and contribute to neuronal dysfunction and death in PD. This is greatly supported by the fact that point mutations and dose-increasing multiplications of the *SNCA* gene, encoding α-synuclein, cause PD [83, 137]. Thus, much of the current efforts to treat PD, by academia and industry alike, are largely concentrated on either preventing the aggregation of α-synuclein [138] or removing its aggregated forms from the patient brain [88]. However, the biological role and clinical significance of α-synuclein aggregation in PD remain elusive. Based on a semiquantitative assessment of α-synuclein positive LP in a series of post-mortem cases, Braak et al. proposed a staging system, comprising six successive stages of LP, starting at the caudal medulla oblongata and gradually ascending to finally become widespread throughout

the neocortex [19]. It was further proposed that the anatomical distribution and load of LP, as described by Braak et al., correlates with (and may explain) the clinical progression and severity of PD [19]. The Braak model of caudorostral spreading of LP has been subsequently challenged, however, by studies showing that only about 50% of PD patients show at pattern of LP consistent with Braak staging [93, 182]. Furthermore, the proposed correlation between Braak LP stage and clinical phenotype has been questioned by multiple studies finding no association between the distribution or load of LP and cognitive decline or extrapyramidal motor dysfunction [135, 90, 179, 126, 34, 89]. Moreover, extensive LP corresponding to Braak stages V-VI can be found in elderly individuals without clinical parkinsonism [19, 110, 95]. In fact, it has been shown that other neuropathological markers such as neurofibrillary tangles and beta-amyloid plaques may correlate better with the incidence and severity of dementia in PD [90, 72, 24]. Another important source of uncertainty regarding the role of LP is the lack of definite correlation between the distribution or load of LP and neuronal loss, both during early and late phases of PD [159, 71].

An increasing body of evidence suggests that impaired protein degradation by the ubiquitin-proteasome system (UPS) and the autophagy-lysosomal pathway play a role in PD. The UPS system is responsible for the degradation of unfolded or misfolded proteins [186], including the removal of α-synuclein [158]. Failure of the UPS is thought to contribute to cellular protein aggregates [187, 32]. Because of their long lifespan, neurons are particularly vulnerable to protein aggregation due to reduced performance of proteasomal function. The UPS also acts outside the cell body in axonal and presynaptic regions to enable time-efficient control of protein levels, for example, to support neurotransmission [154].

Evidence supporting that impaired UPS function may be implicated in the pathogenesis of PD includes 1) the presence of α-synuclein *and* ubiquitin aggregates within Lewy bodies [186, 101]. While experiments in cells have shown that UPS failure contributes to α-synuclein aggregation, new studies suggest a vicious cycle in which α-synuclein aggregates are proposed to actively contribute to UPS impairment [187, 112, 149]. 2) Evidence of quantitative and functional decline of the proteasome has been found in the SNc of patients with idiopathic PD [121]. 3) Mutations in *PRKN* encoding an E3 ubiquitin ligase cause autosomal recessive PD [98].

While most targeted protein degradation occurs through the UPS, protein aggregates, cell organelles such as mitochondria, and other cytoplasmic components are degraded by lysosomes via autophagy [104]. Several lines of evidence suggest that dysfunction of the lysosome-autophagy system may be contributing to the pathogenesis of PD: 1) The accumulation of Lewy pathology suggest there may be decreased clearance of protein aggregates via autophagy. 2) Genetic variation in the *GBA* gene, encoding the lysosomal enzyme beta-glucocerebrosidase, is the strongest known genetic risk factor for

PD [104, 145]. Furthermore, in PD patients without risk variants in *GBA*, the pathologic accumulation of α-synuclein was found to negatively influence *GBA* localization to the lysosomes and consequently contribute to lysosome dysfunction. It has been suggested that this may trigger a vicious cycle, in which impaired lysosomal function increases α-synuclein aggregate accumulation [104, 120].

1.4 Gene expression studies in the Parkinson's disease brain

1.4.1 DGE and associated pathways in PD

Hypothesis-guided expression analyses targeting PD genes (e.g. *SNCA*, *DJ-1*, *PINK1*) have produced variable and partly conflicting findings without shedding much light on the transcriptomic landscape of PD. These were rapidly replaced by genome-wide transcriptomic studies, both using microarray assays and RNA-Seq.
The most recent (2018) review on DGE in PD reported a total of 63 original studies [17]. Of these, 33 were performed on brain tissue, 26 on blood, 3 using cerebrospinal fluid and one was performed on skin tissue [17]. Of the 33 brain tissue studies, 18 were performed on tissue of the SNc and only 8 on tissue extracted from the frontal cortex. Other brain regions included the amygdala, striatum, putamen and occipital cortex. Only two of the brain tissue studies were performed using RNA-Seq [52, 79].
DEGs in PD have shown remarkably poor concordance across different brain regions and between different studies of the same region [17]. Many studies in brain were performed on SNc tissue using microarrays [49, 53, 18, 124, 123, 184, 75, 68]. These studies vary substantially in terms of applied methodology for quality-control and filtering, read-count normalization, and statistical determination of differential expression. Notably, not all studies undertake formal multiple testing correction, and many do not account for essential confounding factors like sex, age, and the pronounced neuronal loss and gliosis characterizing the SNc of individuals with PD. Moreover, a meta-analysis of 13 datasets from 7 studies of different tissues revealed that applying a standardized analysis methodology did not lead to a marked increase in the concordance of DEGs across studies. In fact, not even a single DEG was found in the intersection of all datasets - or even datasets from the same region (SNc) [160].
Despite the low replication rate at the gene-level, enrichment of DEGs in pre-defined functional gene-sets (i.e., pathways) showed higher concordance across studies. The most consistently reported pathway in the SNc was "dopamine metabolism". Since this biological process is specific to dopaminergic (DA) neurons, a cell-type that is selectively

lost in the PD SNc, it is reasonable to assume that this finding most likely reflects altered cell-composition, rather than disease-specific regulation. Therefore, accounting for altered cellular composition is essential when analyzing gene expression data from bulk brain tissue of individuals with PD and other neurodegenerative diseases. Few previous studies have attempted to mitigate the problem of cell-composition in PD by different approaches, including: 1) comparing their results to other neurodegenerative diseases that also involve neuronal loss in SNc [75], selecting subregions of the SNc with mild neuronal loss [18] or estimating cellular composition *in silico* [27].

Another way to circumvent the problem of different cell-composition between CT and PD in the SNc is to isolate individual DA neurons using laser capture microdissection (LCM). Studies applying this methodology reported no change in the expression of genes involved in dopamine metabolism greene2012current, simunovic2009gene, corroborating the hypothesis that those findings were driven by DA cell loss. Interestingly, a study comparing LCM to bulk tissue expression data of SNc tissue reported high expression of glial-specific genes and only low expression levels of neuron-specific genes in bulk brain tissue [55]. While LCM offers a direct comparison of RNA expression in DA neurons between conditions, contamination with non-neuronal RNA is still possible [55]. An additional limitation is the high amount of total RNA required for the experiment. This remains a challenge in PD patients, as only few DA neurons remain at terminal stages of the disease, and these may not be fully representative of the cells that were lost, due to a survivor bias. These findings highlight the difficulty of DGE analysis in PD brain and its functional interpretation concerning cellular differences due to disease pathology.

An alternative brain region to study disease-related gene expression changes in is the prefrontal cortex. It exhibits PD-related pathology, including α-synuclein pathology and complex I deficiency, but shows only mild cellular changes [47, 58, 133]. Reported altered pathways in prefrontal cortex tissue of PD are, among others, neuronal development, glial- and oligodendrocyte-related function, olfactory transduction pathways, as well as mitochondrial and immune function. While the neuronal loss is less pronounced in the prefrontal cortex area, differences in cellular composition between cases and controls are known to exist. Cell-type related pathways like neuronal development, glial and oligodendrocyte function, suggest that the observed altered gene expression profile reflects, at least partly, changes in the cellular composition.

Finally, studies of other brain regions like the striatum have reported changes in pathways such as UPS and oxidative phosphorylation, similar to those suggested by studies of the SNc and prefrontal cortex tissue (Table 1.1). One important limitation of these studies is that, while the striatum is not affected by neuronal loss in PD, it is substantially denervated as a result of the loss of the dopaminergic input from the SNc. The ensuing nigrostriatal denervation is likely to impact measured expression profiles through two mechanisms: 1) loss of the RNA content of the nigrostriatal synapses will confound

the comparison to control samples, and 2) striatal neurons may show changes reflecting altered neurophysiological activity due to altered synaptic input.

Overall, pathways frequently nominated as altered in PD brain include dopamine-metabolism, protein degradation, mitochondrial function (including OXPHOS), synaptic vesicle dynamics and neuroinflammation [17]. A simplified overview of up- and down-regulated pathways that have been proposed as affected by DGE in PD are listed in Table 1.1. Mitochondrial function and protein degradation have been reported as downregulated by most studies, irrespective of tissue or methodology, and have been implicated in PD in other research as described above. Altered pathways related to mitochondria include OXPHOS, oxidative damage response and ATP synthesis. Interestingly, altered and specifically decreased expression of genes encoding components of the mitochondrial respiratory chain was reported as *the* transcriptional hallmark of cellular ageing [60]. This is in line with several experimental analyses in various species and tissues, suggesting a decline in mitochondrial function with ageing [20, 28]. As mentioned above, mitochondrial dysfunction is suggested to contribute to the pathogenesis of neurodegenerative diseases like Alzheimer's disease and PD, although direct causal proof remains elusive [5, 46, 31, 161, 42]. PD has been strongly linked to quantitative and functional deficiency of the mitochondrial respiratory complex [58, 143].

The other consistently reported process associated with transcriptional alterations in PD is protein homeostasis, specifically, protein degradation. DGE studies in PD brain highlight the protein degradation pathway of the UPS mainly as downregulated. However, genes encoding heat shock proteins, which perform chaperone activity related to protein refolding and degradation, have been reported as both down- (LMD, DA neurons [147]) and upregulated (SN, prefrontal cortex and other brain areas [75, 184]). Protein degradation through the autophagy-lysosomal pathway has also been implicated as impaired in PD, as discussed above. Gene expression changes in pathways related to protein degradation via the autophagy-lysosomal pathway have been found in the prefrontal cortex and striatum but are not as consistently reported as the UPS [27].

To conclude, transcriptomic analyses of brain tissue in PD face substantial challenges: 1) So far, most studies have been performed on microarrays, limiting the analysis to a predefined probe set. 2) Studies in post-mortem tissue is challenged by transcript-, cell-, tissue- and sample-dependent variability in RNA degradation, biasing quantification of gene expression. 3) Studies in bulk-tissue suffer from low signal-to-noise ratio due to heterogeneous cell-composition across samples and are biased by disease-specific changes in cell-composition (i.e., neuronal loss and gliosis). 4) Gene expression patterns in post-mortem tissue provide a snapshot of end-stage disease, which is not necessarily informative regarding the processes implicated in neurodegeneration. 5) Methodological and pathological variability contributes to the observed lack of concordance across studies at the gene-level.

To date, no studies have focused on the prefrontal cortex using RNA-Seq to investigate gene expression changes in PD while accounting for cellular composition.

1.4.2 Transcript isoforms in the PD brain

Specific isoform expression profiles in the human brain have been associated with neuronal development and ageing [77] as well as with disease [170], including neurodegeneration [111, 139]. Evidence of altered splicing in PD has been reported by studies targeting specific genes [102]. Four splice variants of *SNCA* were reported to show higher expression in PD frontal cortex compared to healthy controls, although only one of them was significantly overexpressed [13]. Furthermore, two splice variants of *PARK2* were found significantly overexpressed in PD brain [85]. These findings were, however, based on small sample sizes and findings are yet to be replicated. Except for these targeted, hypothesis-based studies, the role of transcript isoforms in PD and in particular the role of DTU remains largely unexplored and no genome-wide studies have been carried out to date.

1.4.3 Protein studies in PD

Protein expression analysis in PD has mostly been focused on proteins that were previously linked to the disease. Few hypothesis-free differential proteome analyses in PD have been performed to date and were limited by the detection power of their methodology [51, 140].
Interestingly, in a recent integrated study of transcriptome and proteome in PD, the authors reported that observed protein alterations in PD brain were mainly in disagreement with the observed changes at the gene level [140].
Unfortunately, so far, limitations and difficulties involving proteomic methodologies force us to continue to study gene expression to identify disease-related mechanisms. However, it is unclear whether the expression level of proteins can be directly inferred from the expression level of RNA and whether disease-related functional consequences can be directly concluded from altered transcription levels. While many have investigated the correlation of transcriptome and proteome and studied the predictive value of the transcriptome to protein level, studies do not necessarily agree [23]. Furthermore, how the coupling of the transcriptome and proteome develops with ageing in the brain [178] and particularly in the PD brain remains largely unknown. Further, alterations in the coupling of transcriptome and proteome could give insight into the altered regulation of protein synthesis and degradation, which is of specific interest in PD.

Pathway	SN	SN LCM	PFC	PCC LCM	STR
Neuron related					
Signal transduction	↓	_	↓	_	_
Dopamine metabolism	↓	_	_	_	_
Vesicle mediated transport	↕	_	_	_	_
Neuro transmission	↓	_	_	_	_
Synaptic transmission	↓	_	_	↑	↑
Axon guidance	↕	_	_	↓	↕
Axonal cytoskelleton	↑	_	↑	_	↑
WNT signaling	↑	_	_	_	_
Neurogenesis, neuron development	_	_	↕	_	_
Synaptic function	_	↓	_	_	↕
Degradation pathways					
Autophagy	_	_	↓	_	↓
Ubiquitin proteasome system	↓	↓	↓	↓	↓
Unfolded protein repsonse	_	_	_	_	_
Unfolded protein binding	↓	↓	_	_	_
Heat shock proteins	↑	↓	↑	_	↑
Immune system					
Inflammation	↑	_	_	↑	↑
Neuroinflammation	↕	_	_	_	_
B cell activity	_	_	_	_	_
Immune response	↑	_	↑	_	_
Mitochondria related					
Krebs/tricarboxylic acid cycle	↓	_	_	_	_
Oxidative stress (metallothioneins (MTs)	↑	_	↑	_	↑
Mitochondrial function	↓	↓	_	↓	_
MRC/ oxidative phosphorylation	↓	_	_	_	↓
Oxidative damage response	↑	_	_	↓	_
Mitochondrial encoded genes	↓	↑	_	_	_
ATP synthesis	_	↓	_	_	↓
Mitochondrial membran translocases	↓	_	_	_	_
Other					
extrinsic apoptosis / regulation of apoptosis	_	↑	_	_	_
cell survival	_	↑	_	_	_
Cytoskeletal structure	↕	_	↓	_	↓
Extracelluar matrix structure	↑	_	_	_	↑
Regulation of oligodendrocyte precursor cells	_	_	↓	_	_
Oligodendrocyte function	↓	_	↓	_	↓
Olfactory receptor function	_	_	↓	_	_
Alternative splicing / splicing regulation	↕	_	_	↕	↕

Table 1.1: **Simplified listing of pathways nominated by DGE studies** Up- and downregulated pathways are indicated by "↑" and "↓" respectively. If both up and downregulation, or just deregulation was reported, it is represented as ↕. "_" indicates empty cells, i.e. no is information given. LCM refers to neuron isolation by laser capture microdissection, (i.e., DA neurons in SN, pyramidal neurons in Posterior cingulate cortex (PCC)). STR stands for striatum, PFC for prefrontal cortex. This listing was generated by summarizing and extracting results from studies reviewed in [17] and is not exhaustive but rather a schematic presentation.

Chapter 2

Aims

This thesis aimed to unveil gene expression signatures that are most likely associated with underlying disease mechanisms by considering i) the cellular complexity of bulk brain tissue ii) the diversity of the transcriptome and iii) the relationship between transcriptome and proteome.

- **Paper I**
  Identify differentially expressed genes and associated functional processes in prefrontal cortex tissue of individuals with PD. Assess the confounding effect of altered cell-composition on differential gene expression analyses.

- **Paper II**
  Identify differential transcript usage events and associated functional processes in prefrontal cortex tissue of individuals with PD. Highlighting the importance of differential transcript usage analysis as a complement to conventional gene expression analysis.

- **Paper III**
  Determine if and how the coupling between transcriptome and proteome in the human brain is altered with ageing and in individuals with PD. Identify distinct mRNA-protein correlation patterns in each group and nominate associated biological processes.

Chapter 3

Material and Methods

This chapter describes the material and methods applied in study I, II and III in detail. These descriptions are based on the method section of the respective papers (published for study I and II, in manuscript for study III). The methods were extracted from the articles and adapted and rearranged to allow for a chronological order. While study I and II were based on the material of two cohorts referred to as PW and NBB (described below), these were respectively termed *discovery cohort* and *replication cohort* in the published article of study II. All three studies involved control samples from the PW cohort which we abbreviated with "CT". Although in manuscript III the control group was abbreviated with "HA", we refer to it as CT in the following method description.

3.1 Material

3.1.1 Subject cohorts: Paper I, II and III

The three studies were performed on data derived from fresh-frozen brain tissue. All experiments were conducted in fresh-frozen prefrontal cortex (Brodmann area 9) from a total of 53 individuals from two independent cohorts (PW and NBB) and Norwegian sudden infant death syndrome (SIDS) individuals. Paper I and II included the analysis of the two independent cohorts, PW and NBB, while paper III included the analysis of one of the PW cohort and $N = 4$ SIDS infants (YG). Paper I additionally included an independent cohort (PA) for which published data was reanalyzed.
The PW cohort ($N = 29$) comprised individuals with idiopathic PD ($N = 17$) from the Park-West study, a prospective population-based cohort which has been described in detail [3] and neurologically healthy controls (CT, $N = 11$). Whole-exome sequencing

had been performed on all patients [62] and known/predicted pathogenic mutations in genes implicated in Mendelian PD and other monogenic neurological disorders had been excluded. None of the study participants had clinical signs of mitochondrial disease or use of medication known to influence mitochondrial function. Controls had no known neurological disease and were matched for age and gender. The NBB cohort consists of $N = 21$ samples from the Netherlands Brain Bank, including idiopathic PD ($N = 10$) and demographically matched neurologically healthy controls ($N = 11$). Individuals with PD fulfilled the National Institute of Neurological Disorders and Stroke [64] and the UK Parkinson's disease Society Brain Bank [176] diagnostic criteria for the disease at their final visit. Ethical permission for these studies was obtained from our regional ethics committee (REK 2017/2082, 2010/1700, 131.04).
To investigate the effect of the rRNA depletion and random primer capture protocol compared to the prevailing poly-A method, we re-analyzed an RNA-seq dataset from a previous publication which employed a poly-A tail selection kit on post-mortem tissue of the same brain area and same disease (PA cohort, $N = 29$ PD samples, $N = 44$ neurologically healthy controls, all males; GEO: GSE68719) [51]. Informed consent was available from all individuals.

3.1.2 Tissue collection and neuropathology: Paper I, II and III

Brains were collected at autopsy and split sagittally along the corpus callosum. One hemisphere was fixed whole in formaldehyde and the other coronally sectioned and snap-frozen in liquid nitrogen. All samples were collected using a standard technique and fixation time of ~ 2 weeks. There was no significant difference in post-mortem interval (PMI) (independent t-test, PW cohort $p = 0.16$; NBB cohort $p = 0.92$), age (independent t-test, PW cohort $p = 0.18$; NBB cohort $p = 0.074$) or gender (independent t-test, PW cohort $p = 0.94$; NBB cohort $p = 0.53$) between PD subjects and controls. Routine neuropathological examination including immunohistochemistry for α-synuclein, tau and beta-amyloid was performed on all brains. All cases showed neuropathological changes consistent with PD including degeneration of the dopaminergic neurons of the SNc in the presence of Lewy pathology. No pathological evidence of neurodegeneration or other neurological disease was found in CT and infants.

3.2 Experimental methods

3.2.1 RNA sequencing: Paper I, II and III

Total RNA was extracted from prefrontal cortex tissue homogenate for all samples using RNeasy plus mini kit (Qiagen) with on-column DNase treatment according to manufacturer's protocol. Final elution was made in 65 μl of dH2O. The concentration and integrity of the total RNA were estimated by Ribogreen assay (Thermo Fisher Scientific), and Fragment Analyzer (Advanced Analytical), respectively. Five hundred ng of total RNA was required for proceeding to downstream RNA-seq applications. First, ribosomal RNA (rRNA) was removed using Ribo-Zero™ Gold (Epidemiology) kit (Illumina, San Diego, CA) using manufacturer's recommended protocol. Immediately after the rRNA removal the RNA was fragmented and primed for the first strand synthesis using the NEBNext First Strand synthesis module (New England BioLabs Inc., Ipswich, MA). Directional second strand synthesis was performed using NEBNext Ultra Directional second strand synthesis kit. Following this, the samples were taken into standard library preparation protocol using NEBNext DNA Library Prep Master Mix Set for Illumina with slight modifications. Briefly, end-repair was done followed by poly-A addition and custom adapter ligation. Post-ligated materials were individually barcoded with unique in-house Genomic Services Lab (GSL) primers and amplified through 12 cycles of PCR. Library quantity was assessed by Picogreen Assay (Thermo Fisher Scientific), and the library quality was estimated by utilizing a DNA High Sense chip on a Caliper Gx (Perkin Elmer). Accurate quantification of the final libraries for sequencing applications was determined using the qPCR-based KAPA Biosystems Library Quantification kit (Kapa Biosystems, Inc.). Each library was diluted to a final concentration of 12.5 nM and pooled equimolar prior to clustering. 125 bp Paired-End (PE) sequencing was performed on an Illumina HiSeq2500 sequencer (Illumina, Inc.) at a target depth of 60 million reads per sample.

In study I and II we tested for differences in RNA quality, measured by the RNA integrity number (RIN) and found that it varied across samples ($mean = 5.3$, $range = 3.0 - 7.2$ for PW; $mean = 6.8$, $range = 3.2 - 9.1$ for NBB), although the difference between conditions did not reach statistical significance in any of the cohorts (t-test $p = 0.72$ and $p = 0.90$ for PW and NBB cohorts, respectively).

For study III we employed the DV200 score and found RNA quality varied across samples ($median_{YG} = 92$, $median_{CT} = 88$, $median_{PD} = 89$), although the difference between groups was not statistically significant ($p_{YG,CT} = 0.06$, $p_{CT,PD} = 0.74$, $p_{YG,PD} = 0.07$, Wilcoxon rank sum test).

3.2.2 qPCR analysis for confirmation of DTU events: paper II

RNA extraction was carried out using the RNeasy Lipid Tissue Mini Kit (QIAGEN 74804), starting with approximately 20 mg brain tissue from three individuals with PD and three controls. 500 ng total RNA were subjected to cDNA synthesis using the SuperScript IV VILO Master Mix with ezDNase Enzyme (Thermofisher Scientific 11766500). Experiments were carried out in triplicates starting with a new cDNA synthesis from aliquoted total RNA. For the SYBR Green quantitative PCR analysis, the PowerUp SYBR Green Master Mix (Thermofisher Scientific, A25776) was used with a thermal cycling of one cycle at 95°C for 20s and 40 cycles at 95°C for 3s and 60°C for 30s on a StepOnePlus instrument (Thermofisher Scientific), and with the primers listed in Table 5.

3.2.3 Proteomics: Paper III

Lysis and protein digestion

$10\mu L$ of lysis buffer (4% SDS, $0.01M$ TRIS pH 7.6) was added to $1mg$ of brain tissue. The tissue was mechanically lysed using Precellys CK 14 ceramic beads, together with the Precellys Evolution (Bertin Corp, Rockville MD, USA). Lysed tissue was transferred to Eppendorf tubes and heated to 95°C for 5 minutes, before centrifugation at $10.000g$ for 5 minutes. The clarified supernatant was transferred to new Eppendorf tubes. Protein measurement was performed using the Pierce BCA protein assay kit (Thermo Fisher). The samples were mixed with up to $50\mu L$ of the clarified lysate with $200\mu L$ of 8 M urea in 0.1 M Tris/HCl pH 8.5 in the filter unit (Microcon YM-30 (Millipore, Cat. MRCF0R030)) and centrifuged at 14,000 × g for 30 min and repeated twice. In total $30\mu g$ of protein per sample was used. The samples were reduced with $10mM$ DTT (1h, RT) and alkylated using $50mM$ IAA (1h, RT), and digested overnight at 37°C with 1:50 enzyme: substrate ratio of sequencing grade trypsin (Promega, Madison, WI). Following digestion, samples were acidified with formic acid and desalted using HLB Oasis SPE cartridges (Waters, Milford, MA). Samples were eluted with 80% acetonitrile in 0.1% formic acid and lyophilized. Peptides were stored at −80°C until use [81].

TMT labeling and fractionation

Digested peptides from each sample were chemically labelled with TMT reagents 10 plex (Thermo Fisher). Peptides were resuspended in a $30\mu L$ resuspension buffer containing

$0.1M$ TEAB (Triethylammonium bicarbonate). TMT reagents ($0.1mg$) were dissolved in $41\mu L$ of anhydrous ACN of which $20\mu L$ was added to the peptides. Following incubation at RT for 1 h, the reaction was quenched using 5% hydroxylamine in HEPES buffer for 15 min at RT. The TMT-labeled samples were pooled at equal protein ratios followed by vacuum centrifuge to near dryness and desalting using Oasis PRIME HLB cartridges. Peptides were fractionated into 8 fractions using the Pierce High pH Reverse-phase Peptide fractionation kit (Thermo Fisher Scientific). The TMT experiment batch setup included additional samples which were not considered in the analysis but included in the preprocessing (filtering and normalization) of the proteomics data.

Liquid Chromatography and Mass Spectrometry Analysis

Each sample was freeze-dried in a Centrivap Concentrator (Labconco) and dissolved in 2% ACN, 1% FA. Approximately 0.5 μg of peptides from each fraction was injected into an Ultimate 3000 RSLC system (Thermo Scientific) connected to a Q-Exactive HF equipped with an EASY-spray ion source (Thermo Scientific). The samples were loaded and desalted on a precolumn (Acclaim PepMap 100, 2 cm $\cdot 75\mu m$ i.d. nanoViper column, packed with 3 μm C18 beads) at a flow rate of $3\frac{\mu L}{min}$ for 5 min with 0.1% TFA. The peptides were separated during a biphasic ACN gradient from two nanoflow UPLC pumps (flow rate of $0.200\frac{\mu L}{min}$) on a $50cm$ analytical column (PepMap RSLC, 50 cm $\cdot 75\mu m$ i.d. EASY-spray column, packed with $2\mu m$ C18 beads (Thermo Scientific). Solvent A was 0.1% FA in water, and Solvent B was 100% ACN. The mass spectrometer was operated in data-dependent acquisition mode to automatically switch between full scan MS1 and MS2 acquisition. The instrument was controlled through Q Excative HF Tune 2.4 and Xcalibur 3.0. MS spectra were acquired in the scan range of $375-1500$ m/z with resolution of $60,000$ at m/z 200, automatic gain control (AGC) target of $3\cdot 10^6$, and a maximum injection time (IT) of $50ms$. The 12 most intense eluting peptides above intensity threshold $6\cdot 10^4$, and charge states two or higher, were sequentially isolated for higher energy collision dissociation (HCD) fragmentation and MS2 acquisition to a normalized HCD collision energy of 32%, target AGC value of $1\cdot 10^5$, resolution $R = 60,000$, and IT of 110 ms. The precursor isolation window was set to $1.6m/z$ with an isolation offset of 0.3 and a dynamic exclusion of $30s$. Lock-mass (445.12003 m/z) internal calibration was used, and isotope exclusion was active.
Raw data were analyzed by MaxQuant v1.5.5.1 [36] with "Variable Modifications" set for TMT 10-plex 126, 127N, 127C, 128N 128C, 129N, 129C, 130N, 130C, 131 to be at N-termini, as well as lysine for database searching and peptide identification.

3.3 Analytical methods

3.3.1 Transcript count estimation: Paper I, II and III

Data quality control

FASTQ files were trimmed using Trimmomatic version 0.36 [16] to remove potential Illumina adapters and low-quality bases with the following parameters: ILLUMINACLIP:truseq.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15. FASTQ files were assessed using fastQC version 0.11.5 [6] prior and following trimming. For an in-depth quality assessment in study I, we mapped the trimmed reads using HISAT2 version 2.1.0 [97] against the hg19 human reference genome (using –rna-strandness RF option) preserving lane-specific information. To discard potential lane-specific sequencing batch effects we inspected the output of the CollectRnaSeqMetrics tool of Picard Tools version 2.6 [86]. Mapping efficiency and proportion of reads mapping to rRNA, intronic, intergenic and coding regions were obtained from the output of the CollectRnaSeqMetrics.
For the poly-A capture dataset [51] in study I, raw FASTQ files were obtained from the Gene Expression Omnibus (GEO:GSE68719) and analyzed exactly as described for our cohorts (with the exception of –rna-strandness in HISAT2, which was turned off to take into account that the cDNA library of this cohort was unstranded).

RNA expression quantification

We used Salmon version 0.9.1 [136] to quantify the abundance at the transcript level with the fragment-level GC bias correction option (–gcBias) and the appropriate option for the library type (–l ISR) against the Ensembl release 75 transcriptome.
For study I, transcript-level quantification was aggregated to gene-level counts using the tximport R package version 1.8.0 [152] using the gene annotations provided by the same Ensembl release (v75).
For study II, we excluded X and Y chromosomes from the GRCh37 reference genome, restricting quantification to transcripts located on autosomes. Quantification obtained from Salmon were scaled using the R package tximport [152] with the scaling method `scaledTPM`, the favoured scaling method for DTU [115].
For study III, we used an updated version of Salmon (1.3.0) to quantify the abundance at the transcript level with the fragment-level GC bias correction option (–gcBias) using the updated GENCODE Release 32 (GRCh38.p13) reference transcriptome and the GRCh38 reference genome, included as decoy [155]. Transcript counts were col-

lapsed to gene-level using R package tximport version 1.14.2 with default parameters (i.e. `countsFromAbundances = FALSE`) and the GENCODE Release 32 (GRCh38.p13) annotation.

3.3.2 Transcript and gene pre-filtering

Paper I

We filtered out genes in non-canonical chromosomes and scaffolds and transcripts encoded by the mitochondrial genome. To further reduce the potential for artefacts we filtered out transcripts with unusually high expression by removing transcripts that gathered more than 1% of the reads on more than half of the samples, which resulted in the removal of 3 and 4 transcripts from the PW and NBB cohorts, respectively. Additionally, low-expressed (i.e., genes whose expression was below the median expression in at least 20% of the samples) were filtered out from downstream analyses. Samples were then marked as outliers if their median correlation in gene expression (log counts per million) with the other samples was below $Q1 - 1.5 \cdot IQR$ or above $Q3 + 1.5 \cdot IQR$ (Tukey's fences; Q1: first quartile, Q3: third quartile, IQR: inter-quartile range). As a result, 3 samples were marked as outliers in the PW cohort and 3 in the NBB cohort, and were not included in downstream analyses (resulting sample sizes: $PW = 26$, $NBB = 18$).

Paper II

Due to the complexity of the human transcriptome in terms of diversity and number of transcripts per gene, DTU methodologies tend to exhibit a worse performance considering the false discovery rate (FDR) when compared to simpler organisms [153]. However, FDR can be reduced considerably if the collection of transcripts undergoes filtering prior to analysis [153]. Transcript filtering, in addition, alleviates the DRIMSeq's (the tool employed for DTU analysis) difficulty to capture the full bandwidth of transcript dispersion through the common gene-level dispersion estimate [128], which results otherwise in a decrease in performance for genes with increasing number of transcripts. We thus excluded low expressed transcripts with a soft filter, allowing for a certain percentage of all samples to have a transcript expression below the given threshold. This filtering methodology was chosen over hard filtering in order to avoid overlooking cases of DTU driven by lack of expression in one of the groups being compared. Using the filtering method available in the DRIMSeq package, we excluded transcripts for which more than `N = min(#Controls, #PD)` samples did not reach 10 read counts or for which their relative

contribution to the overall gene expression was smaller than one percent. In addition, we filtered out genes with less than 10 counts in any one sample. To investigate changes in transcript usage between PD and controls, the resulting filtered set of transcript-level counts were used as an input for both DEXSeq and DRIMSeq as recently suggested by [115]. Analyses were carried out independently on both cohorts.

Paper III

Genes were filtered out if unusually highly expressed (i.e., if they accounted for more than 1% of a sample's library size in more than 50% of all the neurologically healthy samples (i.e. YG and CT). We calculated log_2 transformed counts per million (CPM) for the pre-filtered set of genes. Low-expressed genes $log_2(CPM) < 0.1$, in at least 80% of the samples) were also filtered out. The pre-filtered transcriptomic dataset resulted in a total of $N = 29,6014$ genes. The dataset corresponding to the PD samples, added subsequently in the analyses, was filtered independently following the same filtering approaches and resulting in a total of $N = 29,363$ genes.

3.3.3 Estimation of marker gene profiles: Paper I and II

It has been previously shown that cell-type-specific transcriptional signature patterns derived from bulk tissue samples (marker gene profiles, MGPs), can be used as surrogates for relative cell-type abundance across samples [116]. MGPs for each cell-type are calculated individually across samples, by summarizing the concordant change in their respective marker genes via the first principal component of their expression (i.e., log-transformed counts per million (CPMs)). For the purpose of our study, we calculated MGPs for the main cortical cell-types (neurons, oligodendroglia, microglia, endothelial cells, and astrocytes). Cortical cell-type markers were obtained from the NeuroExpresso database [116], a comprehensive database compiled using mouse brain cell-type expression datasets, and human orthologs were defined using HomoloGene [118]. To reduce the impact of outlier samples, principal component analysis was repeated 100 times on subsampled data, containing an equal number of subjects per group, and removing markers with opposite sign to the main trend. The median score for each sample was used as MGP for the downstream analyses. MGPs obtained with Neuroexpresso-based markers were highly correlated with MGPs calculated using two independent sets of markers from human brain single-cell transcriptomic studies [96, 167]. To assess potential variations associated with the disease across the neuronal markers, we examined the overlap between the markers and the differentially expressed genes in four publicly avail-

able datasets of laser microdissected neurons from PD brain (SNc dopaminergic neurons [26, 55, 148] and posterior cingulate cortex pyramidal neurons [156]). We found minimal overlap (3/78 genes) between our neuronal markers and genes differentially expressed in PD dopaminergic neurons. Moreover, none of the markers were differentially expressed in PD cortical neurons [156]. The vast majority of the cell-type markers used for the calculation of MGPs changed in the same direction across our samples, indicating that MGPs truly represent changes in global cell-type-specific transcription profiles, rather than being driven by changes in specific genes.
In paper I, to unravel potential complex interactions between MGPs and other experimental covariates, including disease status, we calculated the pairwise correlation between all the variables and also their association with the main axes of variation of gene expression. To assist us in choosing an optimal set of MGPs to include as covariates, we quantified the group differences in the cellular proportions between PD and controls using linear models adjusting for the known experimental covariates (i.e., RIN, PMI, sex, age, and sequencing batch). Significant association with disease status was found for oligodendrocyte MGP in the PW cohort and for microglia in the NBB cohort. Thus, these were included in the downstream analyses of study I and II.

3.3.4 Differential analysis

Paper I

We performed differential gene expression analyses using the DESeq2 R package version 1.22.2 [114] with default parameters. Experimental covariates (sex, age, RIN, PMI, and sequencing batch), as well as oligodendrocyte and microglia MGPs, were incorporated into the statistical model. Multiple hypothesis testing was performed with the default automatic filtering of DESeq2 followed by false discovery rate (FDR) calculation by the Benjamini-Hochberg procedure. Analyses were carried out independently for the two cohorts.

Paper II

We performed DTU analysis between PD and controls using two alternative approaches implemented in the tools DRIMSeq [128] and DEXSeq [4]. While DEXSeq was designed for detecting differential exon usage, it is also suitable for assessing DTU by using estimated transcript abundances directly [153, 113, 4]. DRIMSeq was developed specifically for DTU analyses and is based on estimated transcript counts [128]. These

methods assess alternative transcript regulation by directly identifying transcripts that are differentially used, rather than detecting specific splice events. Both methods have shown comparable performance in benchmarks with simulated data [115, 128, 4]. A further advantage was that these tools allow for the inclusion of known covariates into the model design. DRIMSeq assumes a Dirichlet multinomial model for each gene and estimates a gene-wise precision parameter, whereas DEXSeq assumes a negative binomial distribution for counts of each transcript and estimates a transcript-wise dispersion parameter [115]. It is worth noting that DRIMSeq bases its analysis directly on the calculated transcript proportions, thereby modelling the correlation among transcripts in their parent-gene directly, whereas those correlations may not be accurately captured by DEXSeq, as it models each transcript separately and accounts for gene-transcript interaction with a covariate in its model design [115].

Sources of variation in our data were identified using principal component analysis (PCA) at the gene-level. RIN correlated highly with the first principal component, indicating that RNA quality represents a major source of variation in the expression data.

To explore the effect of accounting for disease-associated MGPs in the DTU results, we compared the two alternative designs, with and without oligodendrocyte and microglia MGPs. Accounting for cellular composition slightly increased the discovery signal, identifying a few more DTU genes with both DRIMSeq and DEXSeq. This effect was minor, however, as most DTU genes and events were identified irrespective of whether cell-type composition was accounted for or not.

The results of the DTU analyses were further processed with StageR [165]. Gene-level aggregated p-values (q-values), as well as transcript-level p-values, were passed to stageR for a two-stage screening of significance. For DEXSeq, nominal p-values of all transcripts of a gene were aggregated to a q-value and corrected using the function perGeneQvalue. For DRIMSeq, nominal p-values were already reported at the gene-level and further corrected within stageR using the Benjamini-Hochberg (BH) FDR procedure. To control the family-wise error rate (FWER), transcript-level significance was corrected within-gene, if the gene passed the first screening stage of stageR, with respect to the FDR controlled gene-level significance (q-value). Transcripts of genes that did not pass the first screening stage, were not further assessed for significance at the transcript-level. Nominal transcript-level p-values of both tools were adjusted within StageR using an adapted Holm-Shaffer FWER correction method specifically designed for DTU analysis [165].

We define a transcript as a DTU event, if the FWER-controlled $p < \alpha$, with $\alpha = 0.05$. Similarly, we define as DTU gene any gene that exhibits at least one DTU event.

We define $\alpha = 0.05$ for nominal significance.

3.3.5 Protein intensity normalization and filtering: Paper III

Aggregated protein intensities from maxQuant were further processed in a downstream analysis using R. First, proteins labelled as "Reverse", "Potential.contaminant" and "Only.identified.by.site" were removed from the analysis. In addition, proteins were removed if they exhibited at least one zero intensity in a sample. In order to filter out highly-expressed proteins, we selected the top four highest expressed proteins in each sample (which ranged from 3% to 5% of the total expression of a sample). The union set of these (a total of 19 proteins) was then filtered out from every sample.
We considered three possible normalization approaches for protein quantification, i) raw protein intensities, ii) quantile normalization, and iii) batch effect correction [22] followed by root mean square scaling. To assess each of these strategies we explored the association of the first two components of the principal component analysis (PCA) of the protein expression matrix with the batch variable. Raw protein intensities (i) showed a clear clustering of samples which was associated with the batches of the TMT experiment, which was further amplified by quantile normalization (ii). This effect was no longer noticeable when we applied batch correction (iii), as suggested in [22], where we divided protein intensities by the correction factor based on the reference channels in the respective batches, followed by root mean square scaling.
Additionally, we were able to leverage the RNA-Seq data from the same samples to gain insight into the biological validity of the three alternative normalization options by studying the transcriptome-proteome correlation in the neurologically-healthy groups (CT and YG; log2 transformed values for proteins, and log2 transcript CPMs). The transcriptome-proteome correlation was significantly higher in the batch corrected strategy both across samples and across genes. Based on these observations we chose to apply the batch correction and subsequent root mean square scaling (iii). The pre-filtered proteomic dataset was composed of a total of $N = 2,961$ proteins.

3.3.6 RNA - protein integration: Paper III

We used sparse partial least square (sPLS) as implemented in the mixOmics R package version 6.10.9 [142, 107] to find the linear combinations of variables (transcripts and proteins) that maximize covariance between the transcriptomic and the proteomic layers. sPLS was performed on the pre-filtered transcriptomic (X) and proteomic (Y) datasets using the "canonical" mode and the parameters $keep_X = 50$ and $keep_Y = 50$ for feature selection.
To investigate changes in the transcriptome-proteome correlation between neurologically-healthy groups YG and CT we performed an additional filtering step on both transcripts

and proteins, aiming at increasing the biological signal-to-noise ratio. Genes were flagged for removal if they satisfied at least one of the following criteria: i) not present in the pre-filtered transcriptome, ii) not present in the pre-filtered proteome, iii) low median transcript expression (below 10% quantile), iv) low transcript variance (below 15% quantile). The removal of flagged genes resulted in an analysis-ready dataset of $N = 2,107$ genes.
The dataset corresponding to the PD samples, employed in a subsequent comparison, was filtered independently, following the same filtering approaches and resulting in a slightly lower number of genes in the final analysis-ready list ($N = 1942$). Gene-wise transcript - protein Pearson correlations were calculated across samples independently for each group (CT, PD, YG) using log_2 transformed CPMs for transcript abundance and log_2 transformed batch-corrected and root mean square scaled protein intensities.
Protein-protein interaction networks were generated using the R package coexnet version 1.8.0 [78], which retrieves information on protein co-expression and experimentally evidenced interaction from STRING [162]. Vertices were clustered using the R package igraph version 1.2.5 [40], and its implemented "edge-betweenness" cluster algorithm.

3.3.7 Functional gene-set enrichment analysis

Paper I

Genes were scored according to their significance by transforming the p-values to account for direction of change. For each gene, the up-regulated score was calculated as

$$S_{up} = \begin{cases} 1 - \frac{p}{2}, & \text{if } LFC < 0 \\ \frac{p}{2}, & \text{otherwise} \end{cases} \tag{3.1}$$

, and the down-regulated score as $S_{down} = 1 - S_{up}$, where LFC corresponds to the log fold change and p to the nominal p-value of the gene. Genes were then tested for enrichment using alternatively $log(S_{up})$ and $log(S_{down})$ scores employing the gene score resampling method implemented in the ermineR package version 1.0.1 [117], an R wrapper package for ermineJ [108] with the complete Gene Ontology (GO) database annotation [8] to obtain lists of up-and down-regulated pathways for each cohort.
In order to characterize the main biological processes affected by the cell-type correction, we scored pathways based on the loss of significance caused by the addition of cellular estimates to the gene expression model. We quantified the difference in the level of significance in the up-and down-regulated enrichment results for each significant pathway as $\Delta = log(p_0) - log(p_{CT})$, where p_{CT} and p_0 are the corrected enrichment p-values for

the model with cell- types (CT) and without (0), respectively. Only pathways that were significant in either one of the models were analyzed in this manner ($p_0 < 0.05$ or $p_{CT} < 0.05$).

Paper II

To assess the enrichment of DTU genes in predefined functional gene sets (pathways), we employed the enrichment function of the stringDB R package [162]. DTU genes identified in our discovery cohort were used as hits and all genes surviving the filtering step during pre-processing were used as background. Enrichment was tested for pathways defined by the Genome Ontology (GO) [8, 35]. Each of the three GO categories (Biological Process, Molecular Function, Cellular Compartment) was tested separately. To reduce redundancy of the topmost enriched pathways ($FDR < 0.05$), we performed a clustering in each of the three GO categories. Pathways were clustered by iteratively joining nearest neighbours based on pathway similarity, which we defined with Cohen's kappa coefficient (κ). The similarity of newly formed clusters and unvisited neighbours was iteratively recalculated until no two clusters' κ was higher than a chosen threshold of 0.4. Each cluster was given a representative title, chosen from the names of all the pathways in a cluster. The choice of the cluster title depended on the pathway size, pathway significance or chosen randomly if none of the previous criteria was sufficient. Finally, each pathway cluster was assigned a p-value by aggregating p-values of all cluster members with the Fisher method.
For specific cases of isoform switches between protein-coding transcripts, we used the tool DeepLoc [1] to predict subcellular localization by retrieving the encoded amino acid sequence from the Ensembl release 75.

Paper III

To investigate changes in the transcriptome-proteome correlation between groups, we applied different gene scoring strategies to rank genes according to their change in correlation (δr). For example, to investigate changes occurring in the healthy aging process (i.e., comparing YG vs CT) each gene would be scored by $\delta r = r_{CT} - r_{YG}$. Correspondingly, to investigate changes occurring in the process of ageing with Parkinson's disease, gene scores would be calculated as $\delta r = r_{PD} - r_{YG}$. Finally, changes in transcript-protein correlations between CT and PD groups would be calculated as $\delta r = r_{PD} - r_{CT}$. For each of these three group comparisons (YG→CT, YG→PD, CT→PD), we wanted to specifically identify genes belonging to three functional scenarios in regard to their transcript-protein coupling: a) *"decoupling"*, genes that show a positive transcript-protein correlation in

the reference group (e.g., YG) and loose this correlation ($r \sim 0$) in the other group (e.g., CT); b) *"increased inverse correlation"*, genes which show a correlation above or equal to zero in the reference group and a negative correlation in the other group; and c) *"increased positive correlation"*, genes with a correlation above or equal to zero in the reference group that show an increased positive correlation in the group compared. To this end, gene-specific scores were calculated as follows:

$$\forall R_{ref} > 0 \tag{3.2}$$

$$S_a^i = -|R_{ageing}| + R_{ref} \tag{3.3}$$

$$S_b^i = -R_{ageing} + t(R_{ref}) \tag{3.4}$$

$$S_c^i = R_{ageing} - t(R_{ref}), \tag{3.5}$$

$$with\ t(x) = \frac{x+1}{2} \tag{3.6}$$

, where $i \in 1, 2, 3$ specified the comparison being made:

$$R_{ref} = \begin{cases} R_{YG}, & \text{for i} \in 1, 2 \\ R_{CT}, & \text{for i} = 3 \end{cases} \tag{3.7}$$

$$R_{ageing} = \begin{cases} R_{PD}, & \text{for i} \in 2, 3 \\ R_{CT}, & \text{for i} = 2 \end{cases} \tag{3.8}$$

Heatmaps to visualize scoring distributions were created with the R package ComplexHeatmap [69]. The above gene scorings were used to test for functional enrichment. To this end, we employed the gene score resampling method implemented in the R package ermineR version 1.0.1.9 [117], an R wrapper package for ermineJ [108] with the complete Gene Ontology (GO) database annotation [8] (using aspects: biological process, molecular function and cellular component).

Chapter 4

Summary of results

4.1 Paper I: Common gene expression signatures in Parkinson's disease are driven by changes in cell composition

Multiple studies have examined the transcriptomic signatures associated with PD using post-mortem bulk brain tissue data. While these studies have the potential to shed light on the pathogenesis of PD, they have two major limitations: 1) poor RNA quality, characteristic of post-mortem samples and 2) heterogeneity in cell-type composition, characteristic of brain bulk tissue samples. Here, we carried out the first genome-wide RNA-Seq study employing rRNA depletion and random primer capture (Ribo-Zero approach) in post-mortem brain tissue from neurologically healthy controls (CT) and individuals with idiopathic PD (PD). We studied fresh-frozen prefrontal cortex (Brodmann area 9) from a total of 49 individuals from two independent case-control cohorts, the Norwegian ParkWest study (PW, $N = 29$) and the Netherlands Brain Bank (NBB, $N = 21$). In addition, we reanalyzed data from a previously published genome-wide poly-A RNA-Seq study. To account for the variation in cell-type composition, we used marker gene profiles (MGPs) as surrogates of relative cell-type abundance across the samples.
The main findings from this study were as followed:

1. RNA-Seq data produced with the Ribo-Zero method resulted in more even transcript coverage and substantially less 3'-end bias, compared to poly-A selected RNA-Seq. These findings suggest that the Ribo-Zero approach is superior to poly-A selection, resulting in more accurate mapping and quantification of the transcriptomic data with low RNA quality.

2. RNA quality, represented by the RNA integrity number (RIN), and relative cell type abundance, estimated using MGPs, are the main source of variance in transcriptomic data.

3. Cell composition is confounded with the disease state in PD and has a major impact on the outcome and interpretation of the analyses. We found significant differences between conditions in the relative abundance of oligodendrocytes and microglia. Adjusting for cellular composition by including MGPs for cell-types significantly different between conditions, resulted in: i) overall fewer differentially expressed genes and ii) altered pathway enrichment results.

Specifically, adjusting for cell-types attenuated the enrichment for pathways related to mitochondria, immunity and neuronal functions, while highlighting pathways related to endoplasmic reticulum, unfolded protein response and lipid/fatty acid oxidation. These findings indicate that differential gene expression signatures in PD bulk brain tissue are highly biased by underlying differences in cell-type composition. Modelling cell-type heterogeneity allows us to unveil transcriptomic signatures that are closer to true regulatory changes in the PD brain and are, therefore, more likely to be associated with underlying disease mechanisms.

4.2 Paper II: Differential transcript usage in the Parkinson's disease brain

While multiple studies assessed DGE in the PD brain, the role of alternate splicing, isoform switches and DTU remained largely unexplored. In this work, we performed the first genome-wide DTU study in PD, using the same RNA-Seq dataset as for Paper I. Implementation of rRNA depletion and random primer capture during data generation gave us access to unbiased transcript information, including both coding (mRNA) and non-coding RNA.
The main findings from this study were as followed:

1. The majority of genes exhibiting DTU in PD are not picked up by conventional DGE analysis. Moreover, even for genes exhibiting altered expression in the DGE analysis, DTU assessment provided vital additional information, in some cases, completely changing the conclusion from the findings. For example, DGE analysis indicated upregulation of *MIA* in PD, which is suggestive of an increase in the protein product of this gene. However, DTU analysis revealed that the protein-coding transcript isoform of this gene was in fact downregulated, and the observed

upregulation of *MIA* in the DGE analysis was driven by an increase in the non-coding transcript variant of this gene. Thus, in this case, DGE analysis alone would lead to a misleading functional interpretation.

2. In total, we identified 19 genes showing DTU in PD, which replicated across both the PW and NBB cohorts. Among them, we observed DTU events with functional consequences, due to usage changes between protein-coding and non-coding transcript variants and/or, through changes between variants encoding proteins isoforms with different subcellular localization. Of special interest is the gene *THEM5*. This gene exhibited two isoforms, only one of which is predicted to localize to mitochondria, where it is involved in mitochondrial fatty acid metabolism. Decreased function of this protein has been shown to lead to abnormal mitochondrial morphology and impaired mitochondrial respiration, both of which have been associated with PD. Our analyses indicated down-regulation of the transcript variant encoding the full-length THEM5 protein isoform, predicted to localize to mitochondria, and upregulation of the non-mitochondrial isoform, providing a putative mechanism leading to decreased THEM5 function in PD.

3. Functional gene-set enrichment analyses of DTU events pointed to commonly reported pathways in association with PD, including reactive oxygen species generation and protein homeostasis, suggesting that the identified DTU events are related to molecular mechanisms already associated with the disease.

In conclusion, the findings of this study provided the first insight into the DTU landscape of PD demonstrating that DTU events can have important functional consequences in the PD brain. With this study, we demonstrate that DGE analyses should be complemented by DTU analyses in order to provide a more accurate and complete picture of the functional impact of transcriptomic alterations.

4.3 Paper III: Altered transcriptome-proteome coupling indicates aberrant proteostasis in Parkinson's disease

The correlation between mRNA and protein levels has been shown to decline in the ageing brain, and it has been proposed that this phenomenon may reflect age-dependent changes in proteostasis. It is thought that impaired proteostasis may be implicated in the pathogenesis of PD, but evidence derived from the patient brain is currently limited. Here, we hypothesized that if impaired proteostasis occurs in PD, this should be

reflected in the form of altered correlation between the transcriptome and proteome in the patients' brain, compared to healthy ageing.

To test our hypothesis, we performed transcriptome and proteome-wide analyses in prefrontal cortex tissue from healthy aged individuals (HA, $N = 11$) and PD patients ($N = 17$). To differentiate ageing-related changes from disease-associated alterations in the transcriptome-proteome coupling, we compared correlations to a reference group of 4 infants (YG). Integrating transcriptomics with proteomics provided an additional, highly informative dimension to the gene expression profile of the PD brain.

Comparing gene-wise, across-sample correlations of RNA and protein levels between YG and HA, we observed that most genes decouple (i.e., correlation coefficient r approaches 0) with ageing, in line with previous research. Functional gene-set analysis did not result in significant enrichment for any specific biological processes, suggesting that this decoupling is a widespread, possibly genome-wide phenomenon. Interestingly, we found significant enrichment for synaptic vesicle related processes, in genes that showed inverse/ negative correlation in HA.

The PD brain revealed a more pronounced trend of transcript-protein decoupling than HA, with no significant enrichment of distinct biological pathways, suggesting widespread alterations in the regulation of proteostasis. Interestingly, the PD group was characterized by significant enrichment of mitochondrial respiration and proteasomal pathways among genes with increased positive correlation and increased negative correlation, respectively.

Alterations in proteostasis reflected in negative correlations in PD were observed in genes enriched for proteasomal subunits. We hypothesized that negative correlations across samples can be explained by variable neuronal soma-to-synapse ratios across tissue samples. Spatial separation of mRNA and protein into the neuronal soma and synapses, respectively, could result in negative correlations between mRNA and protein level across samples.

Genes showing inverse correlation in PD were enriched for proteasome subunits, suggesting that these proteins become spatially polarized in PD neurons, with accentuated spatial separation of transcript and protein between the soma and axon/synapses. Moreover, the PD brain was characterized by increased positive mRNA-protein correlation for genes encoding components of the mitochondrial respiratory chain, suggesting these may require tighter regulation in the face of mitochondrial pathology characterizing the PD brain.

Our results are highly consistent with a proteome-wide impairment of proteostasis in the PD brain and strongly support the hypothesis that aberrant proteasomal function is implicated in the pathogenesis of PD. Moreover, our findings have important implications for the correct interpretation of differential gene expression studies in PD. In the presence of a disease-specific altered relationship between transcript and protein levels,

measured differences in mRNA levels cannot be used to confidently predict differences in the encoded proteins and should be supplemented with direct determination of proteins nominated by the analyses.

Chapter 5

Discussion

5.1 Introduction to discussion

Gene expression is the process by which the genotypic information encoded in the genome (i.e., the genotype) determines the phenotype. Typically, information encoded in the DNA is first transcribed to RNA and then translated into protein, i.e., the functional product influencing the phenotype [141]. Genetic and epigenetic disease-related changes ultimately converge on, and influence, gene expression. PD aetiology may be reflected in alterations of both transcriptome and proteome and the identification of these alterations can be linked to molecular mechanisms and functional pathways underlying the disease and further assist in the search for possible therapeutic candidates.
As elaborated in the thesis introduction, previous research of gene expression in PD mainly targeted PD-linked genes. Fewer hypothesis-free transcriptome-wide analyses of differentially expressed genes have been performed. While these helped shaping our understanding of altered gene expression in PD, they also exhibited limitations in both methodology and biological interpretability. The major limitations can be summarized as follows:

1. As most data was generated with microarrays, low expressed genes, and genes not defined in the probe set were missed in the analysis.
2. In analyses of RNA-Seq datasets of post-mortem tissue, RNA degradation was not considered as a possible confounder.
3. Previously favoured brain areas exhibit substantial alterations in cellular composition due to disease pathology confounding observed expression signals.

4. Brain areas exhibiting reduced pathology were employed for gene expression analysis with the assumption that possible altered cellular composition is negligible.

5. The expression landscape at the transcript isoform level has not been adequately explored. Alternatively expressed transcript isoforms were mainly investigated in a hypothesis-guided strategy, targeting PD genes.

6. Functional interpretation of DEGs assumed that these reflect changes of the primary protein-coding transcript isoform, thereby overlooking the possibility of expression changes in transcript isoforms which can encode functionally diverse proteins.

7. Functional interpretation of DEGs assumed a perfect correlation between mRNA and protein level (i.e., changes at the gene level directly infer the same changes at the protein level).

This thesis comprises three complementary approaches aiming to improve our understanding of the molecular changes taking place in the PD brain. By extending conventional gene expression studies, we hoped to surmount the limitations of previous research. Our results revealed several important technical, as well as biological aspects of relevance to the study of PD and, by extension, molecular studies in bulk brain tissue affected by neurodegeneration.

5.2 RNA sequencing using Poly-A enrichment *versus* ribosomal depletion

Until recently, most gene expression studies in PD brain employed microarrays. With the development of RNA-Seq, genome-wide expression studies have become feasible. Two popular approaches to reduce the signal of highly abundant ribosomal RNA were briefly mentioned in section 1.2.1: i) exclusively sequencing polyadenylated mRNA through poly-A enrichment and ii) RNA-Seq after ribosomal depletion. The results of our analyses revealed that ribosomal depletion is advantageous to poly-A when aiming at uncovering disease-associated molecular mechanisms in post-mortem brain tissue.

5.2.1 Poly-A limits gene expression analysis to protein-coding RNA

The most used technique in the few previous PD RNA-Seq studies involved poly-A selection. While poly-A selection is restricted to poly-A transcripts, thereby mostly excluding non-poly-A transcripts such as lncRNA or other non-protein-coding isoforms, ribosomal depletion prior to RNA-Seq enables the quantification of the complete diversity of transcript isoforms. This means that potentially functional non-coding RNAs are included in differential expression analyses, and the complete transcriptomic landscape can be mapped out to assist the exploration of altered transcriptional regulation underlying disease mechanisms. In paper II we reported the occurrence of robust DTU events exhibited by both processed non-coding transcripts and non-protein-coding RNA like lncRNAs. In some cases, these were predicted to have functional consequences at the protein level. With this, we highlighted the necessity to perform transcript isoform analysis in addition to DGE analysis, when aiming to investigate transcriptome-wide expression changes. We concluded that RNA-Seq should not be restricted to poly-A RNA but instead include the sequencing of non-protein-coding RNA, which might be functionally relevant for disease-associated mechanisms. To achieve this, we employed ribosomal depletion prior to RNA-Seq.

5.2.2 Effects of post-mortem degradation can be alleviated by ribosomal depletion

Poly-A enrichment for RNA-Seq involves the hybridization of the poly-A tail with oligo (dT) primers at the mRNA's 3'-end. Comparative analyses of poly-A and ribosomal depleted RNA-Seq datasets investigated the read coverage at 3'- and 5'-ends as well as the overall evenness of coverage across the full length of transcripts. These studies reported that poly-A selection results in 5'- / 3'-end bias, whereby coverage at the 5'-end was reduced, along with a generally non-uniform coverage across the transcript body compared to ribosomal depleted RNA-Seq data [41, 185]. The authors proposed that read coverage in poly-A is dependent on RNA integrity and that partial degradation of transcripts can result in 3'-end bias.

Post-mortem tissue is particularly affected by RNA degradation. This degradation occurs from both 5'- and 3'-ends and introduces substantial biases to expression quantification. In paper I, we re-analyzed poly-A RNA-Seq data of bulk brain tissue and observed increased 3'-end bias along with reduced evenness of read coverage when compared to the ribosomal depleted dataset. We concluded that post-mortem RNA degradation in

combination with poly-A-introduced bias leads to substantially reduced quality of expression quantification.
We conclude that ribosomal depletion prior to RNA-Seq results in reduced 3'-end bias, alleviating the effects of post-mortem degradation and provides, even in post-mortem tissue, near-uniform coverage across the transcript body, thereby reducing bias in expression quantification. Nevertheless, complete removal of RNA degradation induced 3'- and 5'-end bias was not possible.

5.3 Bulk brain tissue complicates expression analyses

5.3.1 Expression variance is correlated with RNA quality

RNA quality is a crucial measure to consider when analyzing RNA of post-mortem tissue samples. RNA quality is assessed prior to sequencing and often represented by measures like RIN. This measure can be informative of the state of degradation in a sample. In study II, we assessed gene expression profiles in reduced dimensionality using principal component (PC) analysis and observed that the first PC correlated with RIN, suggesting that some of the variance in our data is explained by varying degrees of RNA quality. If RIN is confounded with the condition variable it can bias differential expression analyses. These findings stress the importance of adjusting for RNA quality or other measures of RNA degradation, particularly when the studied tissue is susceptible to degradation. We implemented this by adding RIN as covariate in the DGE model design. While this approach effectively reduces noise, difficulties remain.
The rate of RNA degradation in human tissue is not well understood and difficult to account for. Simulated data showed that while many transcripts were highly susceptible to degradation, the degree of degradation varied across genes and cell types [87]. Gene-specific characteristics that were associated with rate of degradation were for example gene length, transcript expression, guanine-cytosine (GC) content [87]. Further, varying degrees of degradation susceptibility across cell-types were observed in blood samples, an effect that most likely also applies to brain tissue with its vast diversity of cell-types[87]. Consequently, an unknown degree of variance due to RNA quality possibly remains, despite adjusting for it by RNA quality measures like RIN.

5.3.2 Cellular composition is reflected in expression data of bulk brain tissue

RNA-Seq studies in bulk brain tissue face the difficulty of noise introduced through cell-type heterogeneity. Differences in cellular composition originate from factors of both biological and technical nature, in the latter case for example during tissue collection. We observed biological variability in cellular composition which included: i) within-group variance and ii) across-group differences.

The heterogeneous disease pathology of PD, including varying degrees of neuronal degradation and microglial infiltration, manifests in variable cell compositions. These are reflected in the gene expression pattern and complicate DGE analyses by decreasing the signal-to-noise ratio.

Differences in cellular composition between conditions confound DGE analyses. In our study, we found that these differences explain much of the observed variance. To mitigate the confounding effect, differences in cellular composition between two conditions need to be accounted for, either by explicitly adjusting for them or by bearing in mind that fold changes could be partly reflecting differences in cellular composition.

Neuronal cell death and the consequential reduced neuronal density in PD patients leads to decreased expression levels of neuron-specific or highly abundant genes within neurons. Thus, functional enrichment of DGE results obtained without formally accounting for cellular composition should be interpreted with consideration for the likely drivers of the observed signal. For example, enrichment for mitochondrial function is at least partly driven by reduced neuronal density in PD tissue, as mitochondrial function-related genes are highly expressed in neuronal cells. The adjustment for cellular composition in the statistical model design then results in an attenuated transcriptional signal of reduced mitochondrial function.

In early transcriptional studies of SNc, differences in neuronal density between conditions were identified as the main contributor to the commonly observed downregulation of dopamine metabolism (introduction section 1.4.1). Following this conclusion, studies increasingly favoured brain areas like the prefrontal cortex (PFC), which are less subject to neurodegeneration. Our findings, however, reveal that even in PFC, disease pathology is reflected in gene expression. The observed varying expression of microglial and oligodendrocyte marker genes gives further evidence of cell composition bias in PFC. While these findings may be specific to PD-associated changes in cellular composition, they are also possibly relevant for other neurodegenerative diseases with pathology that affects cell-type composition. We, therefore, suggest that cellular composition should be considered when studying bulk brain tissue, even when targeting less degenerated brain areas.

5.3.3 Neuronal polarity can influence studies of bulk brain tissue

The architecture or polarization of the neuronal cell-type poses difficulties to both differential expression analyses and integration of transcriptomics and proteomics, particularly in regions comprised of projection neurons, as their axons reach into distal areas [125]. Both tissue sample dissection and biological sample variability can contribute to variance in the soma-to-synapse ratios.
The consequences for differential expression analyses are substantial: i) reduced power, with increase in noise and ii) observed expression variance cannot be confidently assigned to either of the two variables, in cases where soma-to-synapse variance is confounded with conditions. More specifically, comparing tissue samples from conditions that differ in synaptic density can result in associations between decreased expression of synaptic proteins with the condition characterized by the lowest synaptic density. In this case, it is unclear to which extent the observed downregulation can be explained by differences in synapse density or by altered regulation of expression between conditions. This is particularly relevant for neurodegenerative diseases which typically show substantial synaptic loss, as a result of neuronal dysfunction and death.
Similarly, varying soma-to-synapse ratios can generate unexpected signal when integrating transcriptomics with proteomics to study regulatory mechanisms. Both transcript and protein reside throughout the cell in soma, axons and synapses and in cases of spatial separation of the two biological entities, anticorrelating levels of RNA and protein can be observed across samples(/within genes).

5.4 The complexity of transcriptome is not considered in differential gene expression studies

DGE studies are performed with the goal to identify alterations in transcriptional regulation with likely consequences at the protein level. These are assumed to reflect the same change in expression (i.e., increased (decreased) gene expression is interpreted as increased (decreased) expression of the encoded protein). The complexity of the transcriptome and the existence of multiple transcript isoforms per gene are thereby neglected.
Two problems arise from this. First, functionally relevant changes in expression of non-primary transcripts are missed. Secondly, aggregating transcript counts to gene-level counts might produce inaccurate estimates of gene expression, particularly if the se-

quencing method captures mRNA transcripts along with processed non-coding isoforms. One option is to directly study alternative regulation, for example by quantifying reads mapped to splice junctions to identify alternative splice events. However, additional steps are required for functional interpretation, including annotating the event to a gene and further to a transcript isoform and possibly aggregating and quantifying events that were annotated to the same gene.

The development of algorithms that directly estimate the abundance of each isoform simplified the study of alternative transcriptional regulation, for example through DTE or DTU analysis. DTU has the advantage of evaluating transcript isoform expression in relation to the gene's complete transcriptional output, thereby providing a normalized measure to compare isoform expression across groups (see also section 1.2.2).

Additionally, DTU analysis can contribute to a more accurate functional interpretation of DGE results. The modelled transcript usage of each isoform can be compared to the estimated change in expression at the gene level (DGE result) and candidates of transcript isoforms that most likely reflect the gene level change can be narrowed down.

5.5 Functional insights by integrating RNA sequencing data with proteomics

Alterations in gene expression ultimately affect the proteome composition and function. While proteomic studies should, in theory, contain most of the biological relevant signal, they also have several limitations (see also introduction section 1.2.1). For instance, in our dataset, we only identified $\sim 2,900$ proteins, compared to more than $\sim 17,900$ protein-coding genes identified in the RNA-Seq dataset. One informative way to exploit the proteomic dataset is to integrate it with RNA-Seq data, to assess the correlation between RNA and protein level.

5.6 The transcriptional landscape of Parkinson's disease

5.6.1 PD associated alterations in cellular composition reflected in gene expression data

We showed that adjusting for cellular composition dramatically reduces the signal of transcriptional downregulation of mitochondria, a consistently reported signature of PD. While this does not necessarily downplay the involvement of mitochondrial dysfunction in PD [58, 143], it suggests that this dysfunction is not as readily reflected in the transcriptome as previously thought.
One potential explanation for this is a low signal-to-noise ratio in bulk PFC tissue. Pathology studies show that while mitochondrial respiratory defects are expressed in the PFC of PD patients, their distribution is highly variable across neurons, exhibiting a mosaic-pattern where only a few cells show respiratory deficiencies. Thus, it may be difficult to capture this signal in bulk-tissue transcriptomics. Furthermore, impairment of mitochondrial function could be due to other mechanisms acting downstream of transcriptional regulation. The results of study III support this hypothesis, stressing that transcript levels are not always a direct proxy for protein levels, and that regulation of protein synthesis, as well as protein degradation, plays a central role in the regulation of gene expression. Indeed, it was shown that respiratory complex I is proteolytically cleaved upon oxidative stress in cell-models [15]. Finally, our findings indicate that altered cellular composition in PD brain involves changes in microglia and oligodendrocyte density. While evidence for the role of oligodendrocytes in PD pathology is just recently emerging [56], increased microglial differentiation has been suggested by other studies employing complementary methodologies (see Introduction section 1.3.3). We conclude that microglial infiltration also manifests in PFC and oligodendrocytes may play a role in disease pathology.

5.6.2 Altered biological pathways in PD brain

The findings of study I indicate altered expression of genes in biological pathways that have so far not been consistently reported in transcriptomic studies. For example, lipid/fatty acid oxidation was among the top differential gene expression signatures in our studies. Metabolic changes predominantly relating to fatty acid oxidation have been observed in the serum of early-stage PD patients and even suggested as a possible biomarker to diagnose disease onset [25]. However, how this is relevant to the brain

is unknown. In study I, we also discuss that tissue extraction is a possible contributor to differences in cellular composition. Differences in white-grey matter ratio between cases and controls could indeed influence gene expression signatures related to fatty acid metabolism.

Other pathways which were revealed after adjusting for cellular composition suggest a downregulation of endoplasmic reticulum (ER) related function and upregulation of unfolded protein response (UPR). Both ER stress and UPR have been proposed to play a role in neurodegenerative disease pathology [33]. ER stress, in terms of accumulation of unfolded or misfolded proteins in the ER, induces various UPR mechanisms resulting in, for example, modulations of rate of protein synthesis and/or removal of misfolded proteins, either through autophagy or ER-associated protein degradation [82] through the UPS pathway and with the help of heat shock proteins. Finally, when UPR fails to reduce ER stress upon too high abundance of misfolded proteins, apoptosis is triggered [82]. However, the pathological protein aggregation in PD is mainly observed in the cytosol and direct evidence for ER stress or increased UPR is mainly based on in vitro models and not extensively studied in human PD brain [144]. Yet, the frequently reported transcriptional upregulation of heat shock proteins and downregulation of UPS pathways may be further evidence for UPR involvement in disease pathology.

5.6.3 Differential transcript usage analysis indicates altered transcriptional regulation in the PD brain

Robust DTU events which replicated across two independent patient cohorts indicate altered transcriptional regulation the PD brain. Possible contributors include alternative splicing events, alternative transcription start sites as well as alternative cleavage and polyadenylation. The enrichment for biological pathways like reactive oxygen species and ubiquitin-related enzyme activity is in line with commonly reported transcriptional signatures, despite disagreeing fold changes between DTU and DGE results. Interestingly, the most significantly enriched cellular component was the ER, similar to the enrichment for ER-related function in study I. These are important findings for understanding the molecular changes in the PD brain and could help shifting the focus of future studies to transcriptional regulatory processes, or shape hypothesis-guided functional analyses of genes exhibiting PD-associated DTU events.

5.6.4 Altered RNA-protein correlation indicates aberrant proteostasis in PD brain

As summarized in the introduction of this thesis, aberrant proteostasis is believed to be contributing to the pathogenesis of PD. Most of the supporting evidence is, however, either inferred from the observation of protein aggregates in the PD brain [47], based on the study of monogenic diseases [181], whose relevance for idiopathic PD is questionable at best or derived from cell and animal models that do not necessarily reflect human disease [104]. Since PD is, to the best of our knowledge, strictly a human disease, evidence of impaired proteostasis derived from patients is required to confirm this hypothesis. In study III we provide, for the first time, evidence strongly suggesting that proteome-wide aberrant proteostasis occurs in the brain of individuals with PD.
The altered coupling of transcriptome and proteome for some subunits of the proteasome complex adds further support to the importance of the UPS pathway in PD pathology. Based on the negatively correlating RNA and protein levels, we conclude an increased spatial separation between these subunits and their transcript. This lays the ground for future work in which protein abundance in axonal and synaptic regions for these subunits could be investigated to test the hypothesis that PD neurons exhibit alterations in the stoichiometry of their proteasome.

5.7 Stratifying PD samples to reduce noise introduced by disease heterogeneity

In studies I and II, we exploited RNA-Seq data from brain tissue of two independent patient cohorts. This enabled us to perform an independent replication of our findings. We became aware of the low replicability of fold changes at both the level of genes and transcript isoforms. Low concordance of DEGs has been suggested already in previous studies [17]. However, our analyses further highlight the heterogeneity of the disease. We observed great variation in expression across cohorts. While commonly referred to as a single entity, PD exhibits high interindividual variability and diversity. This includes factors like age of onset, constellation and severity of clinical features, rate of progression, response to treatment, risk of complications, type, severity and distribution of underlying pathology [67, 30, 72]. The basis for this clinicopathological diversity remains largely undetermined but is assumed to reflect different, albeit unknown, underlying mechanisms driving the initiation and progression of the disease in different patients [106, 105]. Having no biomarkers stratifying PD according to underlying molecular dysfunction, research is conducted on clinically selected cohorts, which are highly

heterogeneous regarding underlying disease mechanisms and, therefore, also in terms of study readouts. This could be mitigated by stratifying PD samples based on clinical, pathological and molecular features. A successful stratification could purify the expression signal and increase power in DGE studies. Identified molecular signatures could then assist in biomarker detection for distinct subgroups of PD patients.

5.8 Study limitations and caveats

5.8.1 Sample size

While the sample sizes in this study are comparable to or higher than those used in previous research [17], sample size is one of the biggest limitations of this work. As discussed above, disease heterogeneity introduces noise and reduces power in differential expression analyses.
Low sample size is also limiting correlation analyses. In study III, the sample size is particularly small for the infant group ($N = 4$). Nevertheless, we observed high correlation between RNA and protein level, corroborating the previously reported phenomenon of tight coupling between transcriptome and proteome at an early age [178]. Significance of correlation can only be assessed when sample sizes are high and, therefore, we only report general trends of correlation affecting groups of genes enriched in specific functions. To identify more robust correlation patterns, an increased sample size would be beneficial, specifically in the YG group, but also in the CT and PD.

5.8.2 The drawbacks of post-mortem tissue

The studies included in this thesis are founded upon the assumption that transcripts and/or proteins can be accurately identified and quantified in post-mortem brain tissue. Accurate quantification is highly dependent on the quality of RNA, which is subject to post-mortem degradation and therefore negatively associated with post-mortem time interval. While we included RIN as a covariate in the statistical models of both DGE and DTU studies, we are aware that statistically adjusting for RIN cannot fully account for the complex degrees of variance introduced through post-mortem degradation.

5.8.3 Validity of samples as controls

While our control individuals had no clinical history or neuropathological evidence of neurodegenerative disorders, the existence of other neurological disorders or consumption of drugs impacting molecular signatures in the brain cannot be ruled out. In the third paper, we analyzed data from infants with sudden infant death syndrome, for which an underlying brain condition cannot be excluded with confidence.

5.8.4 Tissue collection introduces white-grey matter ratio bias

Brain tissue is comprised of grey and white matter, which are vastly different in terms of cell composition. Grey matter is mainly composed of neuronal cell bodies and most types of glia, whereas white matter contains predominantly myelinated axons, and, to a lesser extent, glial cells, mainly oligodendrocytes [173, 122]. Sample extraction from frozen tissue is performed manually and is therefore subject to human error and variation. Variable white-grey matter ratios across samples could introduce bias if these were group-specific and could further result in gene expression patterns that are confounded by changes in cellular composition.

5.8.5 Integration of RNA sequencing data and proteomics derived from different tissue samples

An important limitation is that the RNA and protein data were derived from the same human donor but not from the exact same tissue sample. While this may result in suboptimal integration of the two omics layers when correlating the RNA and protein readouts, it should not affect the group comparison, which was the focus of our study. Moreover, while not identical, the tissue samples used for RNA and protein were derived from the same area (Brodmann 9) and were immediately adjacent, thus minimizing the discrepancy as much as possible by current technologies.

5.8.6 Disease-associated variation in cellular composition

As already discussed, we found differences in cellular composition to be the main source of variation in our transcriptomic data, specifically differences between conditions. While we did adjust for cellular composition by utilizing known cell type markers, we cannot guarantee to have completely removed cell-type bias due to the following reasons: i) we

have no direct assessment on the validity of the markers ii) adding surrogate variables for cell estimates which are partly confounded with the disease state can bias the effect size estimation of both these variables.
In the DTU analysis, we found that adjusting for cell-types had only minor effects on the results. However, we are aware that an improved adjustment for cellular composition would involve the calculation of cell-type estimates based on known cell-type marker transcript isoforms instead of marker genes. As these are not well established to date, we were not able to apply this.

5.8.7 Adjusting for covariates versus regressing out

As described in the previous section, in study I and II we decided to adjust for known sources of variance in our data (e.g., RIN and cell-type estimates). There are two popular approaches to account for confounding factors in regression analysis that aim to estimate the effect size of variable x_1 on the dependent variable y, with an additional variable x_2, which is confounded with x_1. These are: i) regressing out known variance that can be attributed to x_2 by employing the residuals of the regression of y on x_2 as data to test for the effects of x_1 on y or ii) accounting for confounding factors by including both x_1 and x_2 as covariates into the model design, thereby testing for their effects on y simultaneously [59].
We decided to include covariates in the model design which limited us to the use of statistical tools that allow this (like DESeq2 for DGE and DRIMSeq, DEXSeq for DTU). However, this choice was necessary, as the "regression of residuals" method leads to biased effect estimation of the variable of interest (x_1), which increases with increasing correlation between the independent variables x_1 and x_2, as shown in [59]. This is highly relevant in our analyses where we model expression (y in the toy example) and assume that cell type estimators (x_2) are partly confounded with the condition variable (x_1) [59].

Chapter 6

Conclusions and future directions

6.1 Concluding remarks

The work presented in this thesis reports several novel findings and advances the current knowledge of gene expression changes in PD. The main highlights of our findings are summarized below.

1. Commonly reported gene expression patterns, related to biological processes like mitochondrial function and synaptic transmission, are mainly driven by differences in cell composition.

2. ER-related processes and lipid oxidation are possibly altered and impairment of UPS involved in disease pathology.

3. Expression changes at the gene-level are not conclusive of functional consequences. Often these were not driven by changes in protein-coding transcript isoforms.

4. The PD brain is characterized by profoundly altered correlation between mRNA and protein expression levels. These findings provide robust patient-derived evidence of proteome-wide impairment of proteostasis in the PD brain.

5. Overall, our results stress that caution is needed when interpreting and drawing functional conclusions based on gene expression studies in the brain, in particular for neurodegenerative diseases.

6.2 Future work and outlook

The analysis of bulk brain tissue is challenged by the diversity of cell-types, which varies with age, disease and is influenced by tissue extraction. Data generation is further complicated by complex post-mortem degradation of both RNA and protein. Adjusting for cellular composition with MGPs is one option to mitigate cell-type bias, although we cannot fully assess the validity of these. Cellular deconvolution methods are another option. However, RNA degradation influences the performance of these, as they are often based on gene-expression signatures.

6.2.1 Single-cell RNA sequencing

Our findings highlight the need for single-cell RNA-Seq, which enables the sequencing of RNA of thousands of single nuclei in parallel. While powerful, current single-cell RNA-Seq approaches also have several limitations.

One such limitation is the substantially higher cost in comparison to bulk tissue RNA-Seq. This is particularly critical considering the high heterogeneity of PD, which, as discussed above, becomes apparent in the lack of robust signal in DGE studies. High sample sizes are required and might not be easily affordable.

Secondly, due to post-mortem changes and the complex architecture of brain, single-cell RNA-Seq in patient brain tissue is constrained to nuclei, rather than entire cells. Thus, RNA in neuronal processes and synaptic terminals is not captured. These areas exhibit high abundant and specifically expressed transcripts, which as we proposed in paper III, might be involved in underlying disease mechanisms. Furthermore, the pathological hallmark of aggregated protein in synaptic terminals, in particular α-synuclein, but also other proteins, suggest a key role of synapses in the pathology of PD. Transcriptomic studies, derived from single-cell sequencing which exclude RNA from these areas would therefore potentially overlook relevant functional insight.

Lastly, most single-cell sequencing setups do not involve full-length transcript sequencing but are instead targeting short terminal fragments of the transcript. While single-cell full-length transcript sequencing methodologies have been proposed, they are lacking sensitivity for non-poly-A RNA, due to apparent difficulties in the application of ribosomal depletion [76].

While the rapid methodological improvement might combat these challenges, the problem of missing gene expression signals from synapses persists.

6.2.2 Single-cell proteomics

Finally, to capture consequences of altered transcriptional regulation without inferring protein levels from mRNA levels *and* without cell-type bias, single-cell proteomics seems like the optimal solution that is yet to be fully established [119]. However, also here, only nuclei specific expression would be captured.

6.2.3 Vision

Given the rapid evolution of technologies, computational approaches, and mathematical modelling in the field of neuroscience, one cannot help but be optimistic regarding breakthroughs in the near future. I envision that the gene-expression profile of PD, and other neurodegenerative disorders, will be fully deciphered during my lifetime – most likely via a compound approach integrating bulk-tissue analyses in large sample sizes, with advanced mathematical models to estimate the cell-type composition, and improved single-cell isolation methods allowing us to study all parts of a neuron.

Bibliography

[1] J. J. Almagro Armenteros, C. K. Sønderby, S. K. Sønderby, H. Nielsen, and O. Winther. Deeploc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, 2017.

[2] A. M. Altelaar, J. Munoz, and A. J. Heck. Next-generation proteomics: towards an integrative view of proteome dynamics. *Nature Reviews Genetics*, 14(1):35–48, 2013.

[3] G. Alves, B. Müller, K. Herlofson, I. HogenEsch, W. Telstad, D. Aarsland, O.-B. Tysnes, and J. P. Larsen. Incidence of parkinson's disease in norway: the norwegian parkwest study. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(8):851–857, 2009.

[4] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from rna-seq data. *Nature Precedings*, pages 1–1, 2012.

[5] B. J. Andreone, M. Larhammar, and J. W. Lewcock. Cell death and neurodegeneration. *Cold Spring Harbor Perspectives in Biology*, 12(2):a036434, 2020.

[6] S. Andrews, F. Krueger, A. Segonds-Pichon, L. Biggins, C. Krueger, and S. Wingett. FastQC. Babraham Institute, Jan. 2012.

[7] A. Ascherio and M. A. Schwarzschild. The epidemiology of parkinson's disease: risk factors and prevention. *The Lancet Neurology*, 15(12):1257–1272, 2016.

[8] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[9] B. Aslam, M. Basit, M. A. Nisar, M. Khurshid, and M. H. Rasool. Proteomics: technologies and their applications. *Journal of chromatographic science*, 55(2):182–196, 2017.

[10] W. E. Balch, R. I. Morimoto, A. Dillin, and J. W. Kelly. Adapting proteostasis for disease intervention. *science*, 319(5865):916–919, 2008.

[11] S. P. Barros and S. Offenbacher. Epigenetics: connecting environment and genotype to phenotype and disease. *Journal of dental research*, 88(5):400–408, 2009.

[12] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, K. P. Hall, D. J. Evers, C. L. Barnes, H. R. Bignell, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *nature*, 456(7218):53–59, 2008.

[13] K. Beyer, M. Domingo-Sàbat, J. Humbert, C. Carrato, I. Ferrer, and A. Ariza. Differential expression of alpha-synuclein, parkin, and synphilin-1 isoforms in lewy body disease. *Neurogenetics*, 9(3):163–172, 2008.

[14] C. Blauwendraat, K. Heilbron, C. L. Vallerga, S. Bandres-Ciga, R. von Coelln, L. Pihlstrøm, J. Simón-Sánchez, C. Schulte, M. Sharma, L. Krohn, et al. Parkinson's disease age at onset genome-wide association study: Defining heritability, genetic loci, and α-synuclein mechanisms. *Movement Disorders*, 34(6):866–875, 2019.

[15] J. Blesa, I. Trigo-Damas, A. Quiroga-Varela, and V. R. Jackson-Lewis. Oxidative stress and parkinson's disease. *Frontiers in neuroanatomy*, 9:91, 2015.

[16] A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

[17] G. Borrageiro, W. Haylett, S. Seedat, H. Kuivaniemi, and S. Bardien. A review of genome-wide transcriptomics studies in parkinson's disease. *European Journal of Neuroscience*, 47(1):1–16, 2018.

[18] K. Bossers, G. Meerhoff, R. Balesar, J. W. Van Dongen, C. G. Kruse, D. F. Swaab, and J. Verhaagen. Analysis of gene expression in parkinson's disease: possible involvement of neurotrophic support and axon guidance in dopaminergic cell death. *Brain pathology*, 19(1):91–107, 2009.

[19] H. Braak, K. Del Tredici, U. Rüb, R. A. De Vos, E. N. J. Steur, and E. Braak. Staging of brain pathology related to sporadic parkinson's disease. *Neurobiology of aging*, 24(2):197–211, 2003.

[20] A. Bratic, N.-G. Larsson, et al. The role of mitochondria in aging. *The Journal of clinical investigation*, 123(3):951–957, 2013.

[21] N. L. Bray, H. Pimentel, P. Melsted, and L. Pachter. Near-optimal probabilistic rna-seq quantification. *Nature biotechnology*, 34(5):525–527, 2016.

[22] A. Brenes, J. Hukelmann, D. Bensaddek, and A. I. Lamond. Multibatch tmt reveals false positives, batch effects and missing values. *Molecular & Cellular Proteomics*, 18(10):1967–1980, 2019.

[23] C. Buccitelli and M. Selbach. mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, 2020.

[24] R. E. Burke, W. T. Dauer, and J. P. G. Vonsattel. A critical evaluation of the braak staging scheme for parkinson's disease. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 64(5):485–491, 2008.

[25] F. Burté, D. Houghton, H. Lowes, A. Pyle, S. Nesbitt, A. Yarnall, P. Yu-Wai-Man, D. J. Burn, M. Santibanez-Koref, and G. Hudson. metabolic profiling of parkinson's disease and mild cognitive impairment. *Movement Disorders*, 32(6):927–932, 2017.

[26] I. Cantuti-Castelvetri, C. Keller-McGandy, B. Bouzou, G. Asteris, T. W. Clark, M. P. Frosch, and D. G. Standaert. Effects of gender on nigral gene expression and parkinson disease. *Neurobiology of disease*, 26(3):606–614, 2007.

[27] A. Capurro, L.-G. Bodea, P. Schaefer, R. Luthi-Carter, and V. M. Perreau. Computational deconvolution of genome wide expression data from parkinson's and huntington's disease brain tissues using population-specific expression analysis. *Frontiers in neuroscience*, 8:441, 2015.

[28] A. Cellerino and A. Ori. What have we learned on aging from omics studies? In *Seminars in cell & developmental biology*, volume 70, pages 177–189. Elsevier, 2017.

[29] J. Chen and W. Weiss. Alternative splicing in cancer: implications for biology and therapy. *Oncogene*, 34(1):1–14, 2015.

[30] A. S. Chen-Plotkin, R. Albin, R. Alcalay, D. Babcock, V. Bajaj, D. Bowman, A. Buko, J. Cedarbaum, D. Chelsky, M. R. Cookson, et al. Finding useful biomarkers for parkinson's disease. *Science translational medicine*, 10(454), 2018.

[31] S. J. Chinta, J. K. Mallajosyula, A. Rane, and J. K. Andersen. Mitochondrial alpha-synuclein accumulation impairs complex i function in dopaminergic neurons and results in increased mitophagy in vivo. *Neuroscience letters*, 486(3):235–239, 2010.

[32] A. Ciechanover and Y. T. Kwon. Degradation of misfolded proteins in neurodegenerative diseases: therapeutic targets and strategies. *Experimental & molecular medicine*, 47(3):e147–e147, 2015.

[33] E. Colla. Linking the endoplasmic reticulum to parkinson's disease and alpha-synucleinopathy. *Frontiers in neuroscience*, 13:560, 2019.

[34] C. Colosimo, A. Hughes, L. Kilford, and A. Lees. Lewy body cortical involvement may not always predict dementia in parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(7):852–856, 2003.

[35] G. O. Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.

[36] J. Cox and M. Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.

[37] M. M. Cox and D. L. Nelson. *Lehninger principles of biochemistry.* Wh Freeman, 2008.

[38] F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[39] F. H. Crick. On protein synthesis. In *Symp Soc Exp Biol*, volume 12, page 8, 1958.

[40] M. G. Csardi. Package 'igraph'. *Last accessed*, 3(09):2013, 2013.

[41] P. Cui, Q. Lin, F. Ding, C. Xin, W. Gong, L. Zhang, J. Geng, B. Zhang, X. Yu, J. Yang, et al. A comparison between ribo-minus rna-sequencing and polya-selected rna-sequencing. *Genomics*, 96(5):259–265, 2010.

[42] T. M. Dawson and V. L. Dawson. Mitochondrial mechanisms of neuronal cell death: potential therapeutics. *Annual review of pharmacology and toxicology*, 57:437–454, 2017.

[43] L. M. De Lau and M. M. Breteler. Epidemiology of parkinson's disease. *The Lancet Neurology*, 5(6):525–535, 2006.

[44] M. d. De Rijk, L. Launer, K. Berger, M. Breteler, J. Dartigues, M. Baldereschi, L. Fratiglioni, A. Lobo, J. Martinez-Lage, C. Trenkwalder, et al. Prevalence of parkinson's disease in europe: A collaborative study of population-based cohorts. neurologic diseases in the elderly research group. *Neurology*, 54(11 Suppl 5):S21–3, 2000.

[45] G. P. Delcuve, M. Rastegar, and J. R. Davie. Epigenetic control. *Journal of cellular physiology*, 219(2):243–250, 2009.

[46] L. Devi, V. Raghavendran, B. M. Prabhu, N. G. Avadhani, and H. K. Anandatheerthavarada. Mitochondrial import and accumulation of α-synuclein impair

complex i in human dopaminergic neuronal cultures and parkinson disease brain. *Journal of Biological Chemistry*, 283(14):9089–9100, 2008.

[47] D. W. Dickson. Parkinson's disease and parkinsonism: neuropathology. *Cold Spring Harbor perspectives in medicine*, 2(8):a009258, 2012.

[48] D. W. Dickson. Neuropathology of parkinson disease. *Parkinsonism & related disorders*, 46:S30–S33, 2018.

[49] A. A. Dijkstra, A. Ingrassia, R. X. de Menezes, R. E. van Kesteren, A. J. Rozemuller, P. Heutink, and W. D. van de Berg. Evidence for immune response, axonal dysfunction and reduced endocytosis in the substantia nigra in early stage parkinson's disease. *PloS one*, 10(6):e0128651, 2015.

[50] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. Expression profiling using cdna microarrays. *Nature genetics*, 21(1):10–14, 1999.

[51] A. Dumitriu, J. Golji, A. T. Labadorf, B. Gao, T. G. Beach, R. H. Myers, K. A. Longo, and J. C. Latourelle. Integrative analyses of proteomics and rna transcriptomics implicate mitochondrial processes, protein folding pathways and gwas loci in parkinson disease. *BMC medical genomics*, 9(1):5, 2015.

[52] A. Dumitriu, J. C. Latourelle, T. C. Hadzi, N. Pankratz, D. Garza, J. P. Miller, J. M. Vance, T. Foroud, T. G. Beach, and R. H. Myers. Gene expression profiles in parkinson disease prefrontal cortex implicate foxo1 and genes under its transcriptional regulation. *PLoS Genet*, 8(6):e1002794, 2012.

[53] P. Durrenberger, E. Grünblatt, F. S. Fernando, C. M. Monoranu, J. Evans, P. Riederer, R. Reynolds, D. T. Dexter, et al. Inflammatory pathways in parkinson's disease; a bne microarray study. *Parkinson's disease*, 2012, 2016.

[54] R. Elkon, A. P. Ugalde, and R. Agami. Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, 14(7):496–506, 2013.

[55] M. Elstner, C. M. Morris, K. Heim, A. Bender, D. Mehta, E. Jaros, T. Klopstock, T. Meitinger, D. M. Turnbull, and H. Prokisch. Expression analysis of dopaminergic neurons in parkinson's disease and aging links transcriptional dysregulation of energy metabolism to cell death. *Acta neuropathologica*, 122(1):75, 2011.

[56] O. Errea and M. C. Rodriguez-Oroz. Oligodendrocytes, a new player in the etiology of parkinson's disease. *Movement Disorders*, 36(1):83–83, 2021.

[57] W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microrna: are the answers in sight? *Nature reviews genetics*, 9(2):102–114, 2008.

[58] I. H. Fløtnes, E. Fernandez-Vizarra, M. Lykouri, B. Brakedal, G. O. Skeie, H. Miletic, P. K. Lilleng, G. Alves, O.-B. Tysnes, K. Haugarvoll, et al. Neuronal complex i deficiency occurs throughout the parkinson's disease brain, but is not associated with neurodegeneration or mitochondrial dna damage. *Acta neuropathologica*, 135(3):409–425, 2018.

[59] R. P. Freckleton. On the misuse of residuals in ecology: regression of residuals vs. multiple regression. *Journal of Animal Ecology*, pages 542–545, 2002.

[60] S. Frenk and J. Houseley. Gene expression hallmarks of cellular ageing. *Biogerontology*, 19(6):547–566, 2018.

[61] K. Froussios, K. Mourão, G. Simpson, G. Barton, and N. Schurch. Relative abundance of transcripts (rats): Identifying differential isoform abundance from rna-seq. *F1000Research*, 8, 2019.

[62] J. J. Gaare, G. S. Nido, P. Sztromwasser, P. M. Knappskog, O. Dahl, M. Lund-Johansen, J. Maple-Grødem, G. Alves, O.-B. Tysnes, S. Johansson, et al. Rare genetic variation in mitochondrial pathways influences the risk for parkinson's disease. *Movement Disorders*, 33(10):1591–1600, 2018.

[63] J. J. Gaare, G. O. Skeie, C. Tzoulis, J. P. Larsen, and O.-B. Tysnes. Familial aggregation of parkinson's disease may affect progression of motor symptoms and dementia. *Movement Disorders*, 32(2):241–245, 2017.

[64] D. J. Gelb, E. Oliver, and S. Gilman. Diagnostic criteria for parkinson disease. *Archives of neurology*, 56(1):33–39, 1999.

[65] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss. Rna-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986, 2008.

[66] C. L. Gooch, E. Pracht, and A. R. Borenstein. The burden of neurological disease in the united states: A summary report and call to action. *Annals of neurology*, 81(4):479–484, 2017.

[67] J. C. Greenland, C. H. Williams-Gray, and R. A. Barker. The clinical heterogeneity of parkinson's disease and its therapeutic implications. *European Journal of Neuroscience*, 49(3):328–338, 2019.

[68] E. Grünblatt, S. Mandel, J. Jacob-Hirsch, S. Zeligson, N. Amariglo, G. Rechavi, J. Li, R. Ravid, W. Roggendorf, P. Riederer, et al. Gene expression profiling of parkinsonian substantia nigra pars compacta; alterations in ubiquitin-proteasome, heat shock protein, iron and oxidative stress regulated proteins, cell

adhesion/cellular matrix and vesicle trafficking genes. *Journal of neural transmission*, 111(12):1543–1573, 2004.

[69] Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.

[70] A. Gustavsson, M. Svensson, F. Jacobi, C. Allgulander, J. Alonso, E. Beghi, R. Dodel, M. Ekman, C. Faravelli, L. Fratiglioni, et al. Cost of disorders of the brain in europe 2010. *European neuropsychopharmacology*, 21(10):718–779, 2011.

[71] G. Halliday, K. Del Tredici, and H. Braak. Critical appraisal of brain pathology staging related to presymptomatic and symptomatic cases of sporadic parkinson's disease. *Parkinson's Disease and Related Disorders*, pages 99–103, 2006.

[72] G. M. Halliday and H. McCann. The progression of pathology in parkinson's disease. *Annals of the New York Academy of Sciences*, 1184(1):188–195, 2010.

[73] T. H. Hamza and H. Payami. The heritability of risk and age at onset of parkinson's disease after accounting for known genetic risk factors. *Journal of human genetics*, 55(4):241–243, 2010.

[74] N. Hattori, M. Tanaka, T. Ozawa, and Y. Mizuno. Immunohistochemical studies on complexes i, ii, iii, and iv of mitochondria in parkinson's disease. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 30(4):563–571, 1991.

[75] M. A. Hauser, Y.-J. Li, H. Xu, M. A. Noureddine, Y. S. Shao, S. R. Gullans, C. R. Scherzer, R. V. Jensen, A. C. McLaurin, J. R. Gibson, et al. Expression profiling of substantia nigra in parkinson disease, progressive supranuclear palsy, and frontotemporal dementia with parkinsonism. *Archives of neurology*, 62(6):917–921, 2005.

[76] T. Hayashi, H. Ozaki, Y. Sasagawa, M. Umeda, H. Danno, and I. Nikaido. Single-cell full-length total rna sequencing uncovers dynamics of recursive splicing and enhancer rnas. *Nature communications*, 9(1):1–16, 2018.

[77] M. M. Hefti, K. Farrell, S. Kim, K. R. Bowles, M. E. Fowkes, T. Raj, and J. F. Crary. High-resolution temporal and regional mapping of mapt expression and splicing in human brain development. *PloS one*, 13(4):e0195771, 2018.

[78] J. D. Henao. coexnet: An r package to build co-expression networks from microarray data. 2018.

[79] A. Henderson-Smith, J. J. Corneveaux, M. De Both, L. Cuyugan, W. S. Liang, M. Huentelman, C. Adler, E. Driver-Dunckley, T. G. Beach, and T. L. Dunckley. Next-generation profiling to identify the molecular etiology of parkinson dementia. *Neurology Genetics*, 2(3), 2016.

[80] M. A. Hernán, B. Takkouche, F. Caamaño-Isorna, and J. J. Gestal-Otero. A meta-analysis of coffee drinking, cigarette smoking, and the risk of parkinson's disease. *Annals of neurology*, 52(3):276–284, 2002.

[81] M. Hernandez-Valladares, E. Aasebø, O. Mjaavatten, M. Vaudel, Ø. Bruserud, F. Berven, and F. Selheim. Reliable fasp-based procedures for optimal quantitative proteomic and phosphoproteomic analysis on samples from acute myeloid leukemia patients. *Biological procedures online*, 18(1):13, 2016.

[82] C. Hetz, K. Zhang, and R. J. Kaufman. Mechanisms, regulation and functions of the unfolded protein response. *Nature Reviews Molecular Cell Biology*, 21(8):421–438, 2020.

[83] B. A. Hijaz and L. A. Volpicelli-Daley. Initiation and propagation of α-synuclein aggregation in the nervous system. *Molecular neurodegeneration*, 15(1):1–12, 2020.

[84] M. S. Hipp, P. Kasturi, and F. U. Hartl. The proteostasis network and its decline in ageing. *Nature reviews Molecular cell biology*, 20(7):421–435, 2019.

[85] J. Humbert, K. Beyer, C. Carrato, J. L. Mate, I. Ferrer, and A. Ariza. Parkin and synphilin-1 isoform expression changes in lewy body diseases. *Neurobiology of disease*, 26(3):681–687, 2007.

[86] B. Institue. Picard tools—by broad institute, 2018.

[87] A. E. Jaffe, R. Tao, A. L. Norris, M. Kealhofer, A. Nellore, J. H. Shin, D. Kim, Y. Jia, T. M. Hyde, J. E. Kleinman, et al. qsva framework for rna quality correction in differential expression analysis. *Proceedings of the National Academy of Sciences*, 114(27):7130–7135, 2017.

[88] J. Jankovic, I. Goodman, B. Safirstein, T. K. Marmon, D. B. Schenk, M. Koller, W. Zago, D. K. Ness, S. G. Griffith, M. Grundman, et al. Safety and tolerability of multiple ascending doses of prx002/rg7935, an anti–α-synuclein monoclonal antibody, in patients with parkinson disease: a randomized clinical trial. *JAMA neurology*, 75(10):1206–1214, 2018.

[89] K. Jellinger. Morphological substrates of parkinsonism with and without dementia: a retrospective clinico-pathological study. *Neuropsychiatric Disorders An Integrative Approach*, pages 91–104, 2007.

[90] K. A. Jellinger. A critical evaluation of current staging of α-synuclein pathology in lewy body disorders. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1792(7):730–740, 2009.

[91] W. Jiang and L. Chen. Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing. *Computational and Structural Biotechnology Journal*, 19:183–195, 2021.

[92] W. Johannsen. The genotype conception of heredity. *The American Naturalist*, 45(531):129–159, 1911.

[93] M. Kalaitzakis, M. Graeber, S. Gentleman, and R. Pearce. The dorsal motor nucleus of the vagus is not an obligatory trigger site of parkinson's disease: a critical analysis of α-synuclein staging. *Neuropathology and applied neurobiology*, 34(3):284–295, 2008.

[94] L. Kalia and A. Lang. Parkinson's disease. *The Lancet*, 386(9996):896–912, 2015.

[95] L. Kalia and A. Lang. Parkinson's disease. lancet 386, 896–912. *Molecular Therapy: Methods & Clinical Development*, 2015.

[96] K. W. Kelley, H. Nakao-Inoue, A. V. Molofsky, and M. C. Oldham. Variation among intact tissue samples reveals the core transcriptional features of human cns cell classes. *Nature neuroscience*, 21(9):1171–1184, 2018.

[97] D. Kim, B. Langmead, and S. L. Salzberg. Hisat: a fast spliced aligner with low memory requirements. *Nature methods*, 12(4):357–360, 2015.

[98] 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, 1998.

[99] K. R. Kukurba and S. B. Montgomery. Rna sequencing and analysis. *Cold Spring Harbor Protocols*, 2015(11):pdb–top084970, 2015.

[100] B. T. Kurien and R. H. Scofield. Western blotting. *Methods*, 38(4):283–293, 2006.

[101] S. Kuzuhara, H. Mori, N. Izumiyama, M. Yoshimura, and Y. Ihara. Lewy bodies are ubiquitinated. *Acta neuropathologica*, 75(4):345–353, 1988.

[102] V. La Cognata, V. D'Agata, F. Cavalcanti, and S. Cavallaro. Splicing: is there an alternative contribution to parkinson's disease? *Neurogenetics*, 16(4):245–263, 2015.

[103] K. Lage, N. T. Hansen, E. O. Karlberg, A. C. Eklund, F. S. Roque, P. K. Donahoe, Z. Szallasi, T. S. Jensen, and S. Brunak. A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences*, 105(52):20870–20875, 2008.

[104] J. D. Lane, V. I. Korolchuk, J. T. Murray, C. Karabiyik, M. J. Lee, and D. C. Rubinsztein. Autophagy impairment in parkinson's disease. *Essays in biochemistry*, 61(6):711–720, 2017.

[105] M. Lawton, F. Baig, M. Rolinski, C. Ruffman, K. Nithi, M. T. May, Y. Ben-Shlomo, and M. Hu. Parkinson's disease subtypes in the oxford parkinson disease centre (opdc) discovery cohort. *Journal of Parkinson's disease*, 5(2):269–279, 2015.

[106] M. Lawton, Y. Ben-Shlomo, M. T. May, F. Baig, T. R. Barber, J. C. Klein, D. M. Swallow, N. Malek, K. A. Grosset, N. Bajaj, et al. Developing and validating parkinson's disease subtypes and their motor and cognitive progression. *Journal of Neurology, Neurosurgery & Psychiatry*, 89(12):1279–1287, 2018.

[107] K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

[108] H. K. Lee, W. Braynen, K. Keshav, and P. Pavlidis. Erminej: tool for functional analysis of gene expression data sets. *BMC bioinformatics*, 6(1):269, 2005.

[109] R. M. Lequin. Enzyme immunoassay (eia)/enzyme-linked immunosorbent assay (elisa). *Clinical chemistry*, 51(12):2415–2418, 2005.

[110] S. Lewis, T. Foltynie, A. D. Blackwell, T. W. Robbins, A. M. Owen, and R. A. Barker. Heterogeneity of parkinson's disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(3):343–348, 2005.

[111] L. Lin, J. W. Park, S. Ramachandran, Y. Zhang, Y.-T. Tseng, S. Shen, H. J. Waldvogel, M. A. Curtis, R. L. Faull, J. C. Troncoso, et al. Transcriptome sequencing reveals aberrant alternative splicing in huntington's disease. *Human molecular genetics*, 25(16):3454–3466, 2016.

[112] E. Lindersson, R. Beedholm, P. Højrup, T. Moos, W. Gai, K. B. Hendil, and P. H. Jensen. Proteasomal inhibition by α-synuclein filaments and oligomers. *Journal of Biological Chemistry*, 279(13):12924–12934, 2004.

[113] M. Love, C. Soneson, R. Patro, K. Vitting-seerup, and A. Oshlack. Swimming downstream: statistical analysis of differential transcript usage following salmon

quantification [version 1; peer review: 3 approved with reservations] referee status: This article is included in the rpackage gateway. *Issue May*, 2019.

[114] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

[115] M. I. Love, C. Soneson, and R. Patro. Swimming downstream: statistical analysis of differential transcript usage following salmon quantification. *F1000Research*, 7, 2018.

[116] B. O. Mancarci, L. Toker, S. J. Tripathy, B. Li, B. Rocco, E. Sibille, and P. Pavlidis. Cross-laboratory analysis of brain cell type transcriptomes with applications to interpretation of bulk tissue data. *Eneuro*, 4(6), 2017.

[117] O. Mancarci. erminer: gene set analysis with multifunctionality assessment, 2019. R package version 1.0.1. Available online at:https://github.com/PavlidisLab/ermineR.

[118] O. Mancarci. Homologene: quick access to homologene and gene annotation updates, 2019. R package version 1.4.68. Available online at:https://CRAN.R-project.org/package=homologene.

[119] V. Marx. A dream of single-cell proteomics. *Nature methods*, 16(9):809–812, 2019.

[120] J. R. Mazzulli, Y.-H. Xu, Y. Sun, A. L. Knight, P. J. McLean, G. A. Caldwell, E. Sidransky, G. A. Grabowski, and D. Krainc. Gaucher disease glucocerebrosidase and α-synuclein form a bidirectional pathogenic loop in synucleinopathies. *Cell*, 146(1):37–52, 2011.

[121] K. S. P. McNaught, R. Belizaire, O. Isacson, P. Jenner, and C. W. Olanow. Altered proteasomal function in sporadic parkinson's disease. *Experimental neurology*, 179(1):38–46, 2003.

[122] A. A. Mercadante and P. Tadi. Neuroanatomy, gray matter. In *StatPearls [Internet]*. StatPearls Publishing, 2019.

[123] R. M. Miller, G. L. Kiser, T. Kaysser-Kranich, R. J. Lockner, C. Palaniappan, and H. J. Federoff. Robust dysregulation of gene expression in substantia nigra and striatum in parkinson's disease. *Neurobiology of disease*, 21(2):305–313, 2006.

[124] L. B. Moran, D. Duke, M. Deprez, D. Dexter, R. Pearce, and M. Graeber. Whole genome expression profiling of the medial and lateral substantia nigra in parkinson's disease. *Neurogenetics*, 7(1):1–11, 2006.

[125] C. P. Moritz, T. Mühlhaus, S. Tenzer, T. Schulenborg, and E. Friauf. Poor transcript-protein correlation in the brain: negatively correlating gene products reveal neuronal polarity as a potential cause. *Journal of neurochemistry*, 149(5):582–604, 2019.

[126] C. M. Müller, R. A. de Vos, C.-A. Maurage, D. R. Thal, M. Tolnay, and H. Braak. Staging of sporadic parkinson disease-related α-synuclein pathology: inter-and intra-rater reliability. *Journal of Neuropathology & Experimental Neurology*, 64(7):623–628, 2005.

[127] M. A. Nalls, C. Blauwendraat, C. L. Vallerga, K. Heilbron, S. Bandres-Ciga, D. Chang, M. Tan, D. A. Kia, A. J. Noyce, A. Xue, et al. 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, 2019.

[128] M. Nowicka and M. D. Robinson. Drimseq: a dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*, 5, 2016.

[129] R. L. Nussbaum and C. E. Ellis. Alzheimer's disease and parkinson's disease. *New england journal of medicine*, 348(14):1356–1364, 2003.

[130] S. Ogbourne and T. M. Antalis. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochemical Journal*, 331(1):1–14, 1998.

[131] M. C. Oldham, G. Konopka, K. Iwamoto, P. Langfelder, T. Kato, S. Horvath, and D. H. Geschwind. Functional organization of the transcriptome in human brain. *Nature neuroscience*, 11(11):1271, 2008.

[132] S. Pal and J. K. Tyler. Epigenetics and aging. *Science advances*, 2(7):e1600584, 2016.

[133] W. D. Parker Jr, J. K. Parks, and R. H. Swerdlow. Complex i deficiency in parkinson's disease frontal cortex. *Brain research*, 1189:215–218, 2008.

[134] J. Parkinson. An essay on the shaking palsy (london: Sherwood, neely and jones). 1817.

[135] L. Parkkinen, T. Kauppinen, T. Pirttilä, J. M. Autere, and I. Alafuzoff. α-synuclein pathology does not predict extrapyramidal symptoms or dementia. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 57(1):82–91, 2005.

[136] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419, 2017.

[137] M. H. Polymeropoulos, C. Lavedan, E. Leroy, S. E. Ide, A. Dehejia, A. Dutra, B. Pike, H. Root, J. Rubenstein, R. Boyer, et al. Mutation in the α-synuclein gene identified in families with parkinson's disease. *science*, 276(5321):2045–2047, 1997.

[138] J. Pujols, S. Peña-Díaz, D. F. Lázaro, F. Peccati, F. Pinheiro, D. González, A. Carija, S. Navarro, M. Conde-Giménez, J. García, et al. Small molecule inhibits α-synuclein aggregation, disrupts amyloid fibrils, and prevents degeneration of dopaminergic neurons. *Proceedings of the National Academy of Sciences*, 115(41):10481–10486, 2018.

[139] H. Rhinn, L. Qiang, T. Yamashita, D. Rhee, A. Zolin, W. Vanti, and A. Abeliovich. Alternative α-synuclein transcript usage as a convergent mechanism in parkinson's disease pathology. *Nature communications*, 3(1):1–11, 2012.

[140] B. E. Riley, S. J. Gardai, D. Emig-Agius, M. Bessarabova, A. E. Ivliev, B. Schüle, J. Alexander, W. Wallace, G. M. Halliday, J. W. Langston, et al. Systems-based analyses of brain regions functionally impacted in parkinson's disease reveals underlying causal mechanisms. *PloS one*, 9(8):e102909, 2014.

[141] K. Roberts, B. Alberts, A. Johnson, P. Walter, and T. Hunt. Molecular biology of the cell. *New York: Garland Science*, 2002.

[142] F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752, 2017.

[143] A. Schapira, J. Cooper, D. Dexter, P. Jenner, J. Clark, and C. Marsden. Mitochondrial complex i deficiency in parkinson's disease. *The Lancet*, 333(8649):1269, 1989.

[144] W. Scheper and J. J. Hoozemans. The unfolded protein response in neurodegenerative diseases: a neuropathological perspective. *Acta neuropathologica*, 130(3):315–331, 2015.

[145] D. C. Schöndorf, M. Aureli, F. E. McAllister, C. J. Hindley, F. Mayer, B. Schmid, S. P. Sardi, M. Valsecchi, S. Hoffmann, L. K. Schwarz, et al. ipsc-derived neurons from gba1-associated parkinson's disease patients show autophagic defects and impaired calcium homeostasis. *Nature communications*, 5(1):1–17, 2014.

[146] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature Reviews Genetics*, 15(2):121–132, 2014.

[147] F. Simunovic, M. Yi, Y. Wang, L. Macey, L. T. Brown, A. M. Krichevsky, S. L. Andersen, R. M. Stephens, F. M. Benes, and K. C. Sonntag. Gene expression profiling of substantia nigra dopamine neurons: further insights into parkinson's disease pathology. *Brain*, 132(7):1795–1809, 2009.

[148] F. Simunovic, M. Yi, Y. Wang, R. Stephens, and K. C. Sonntag. Evidence for gender-specific transcriptional profiles of nigral dopamine neurons in parkinson disease. *PloS one*, 5(1):e8856, 2010.

[149] H. Snyder, K. Mensah, C. Theisler, J. Lee, A. Matouschek, and B. Wolozin. Aggregated and monomeric α-synuclein bind to the s6 proteasomal protein and inhibit proteasomal function. *Journal of Biological Chemistry*, 278(14):11753–11759, 2003.

[150] M. Somel, S. Guo, N. Fu, Z. Yan, H. Y. Hu, Y. Xu, Y. Yuan, Z. Ning, Y. Hu, C. Menzel, et al. Microrna, mrna, and protein expression link development and aging in human and macaque brain. *Genome research*, 20(9):1207–1218, 2010.

[151] A. R. Sonawane, J. Platig, M. Fagny, C.-Y. Chen, J. N. Paulson, C. M. Lopes-Ramos, D. L. DeMeo, J. Quackenbush, K. Glass, and M. L. Kuijjer. Understanding tissue-specific gene regulation. *Cell reports*, 21(4):1077–1088, 2017.

[152] C. Soneson, M. I. Love, and M. D. Robinson. Differential analyses for rna-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, 2015.

[153] C. Soneson, K. L. Matthes, M. Nowicka, C. W. Law, and M. D. Robinson. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome biology*, 17(1):1–15, 2016.

[154] S. D. Speese, N. Trotta, C. K. Rodesch, B. Aravamudan, and K. Broadie. The ubiquitin proteasome system acutely regulates presynaptic protein turnover and synaptic efficacy. *Current biology*, 13(11):899–910, 2003.

[155] A. Srivastava, L. Malik, H. Sarkar, M. Zakeri, F. Almodaresi, C. Soneson, M. I. Love, C. Kingsford, and R. Patro. Alignment and mapping methodology influence transcript abundance estimation. *Genome biology*, 21(1):1–29, 2020.

[156] C. Stamper, A. Siegel, W. S. Liang, J. V. Pearson, D. A. Stephan, H. Shill, D. Connor, J. N. Caviness, M. Sabbagh, T. G. Beach, et al. Neuronal gene expression correlates of parkinson's disease with dementia. *Movement disorders: official journal of the Movement Disorder Society*, 23(11):1588–1595, 2008.

[157] R. Stark, M. Grzelak, and J. Hadfield. Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656, 2019.

[158] L. Stefanis, E. Emmanouilidou, M. Pantazopoulou, D. Kirik, K. Vekrellis, and G. K. Tofaris. How is alpha-synuclein cleared from the cell? *Journal of neurochemistry*, 150(5):577–590, 2019.

[159] D. J. Surmeier, J. A. Obeso, and G. M. Halliday. Selective neuronal vulnerability in parkinson disease. *Nature Reviews Neuroscience*, 18(2):101, 2017.

[160] G. T. Sutherland, N. A. Matigian, A. M. Chalk, M. J. Anderson, P. A. Silburn, A. Mackay-Sim, C. A. Wells, and G. D. Mellick. A cross-study transcriptional analysis of parkinson's disease. *PLoS one*, 4(3):e4955, 2009.

[161] R. H. Swerdlow, J. M. Burns, and S. M. Khan. The alzheimer's disease mitochondrial cascade hypothesis. *Journal of Alzheimer's Disease*, 20(s2):S265–S279, 2010.

[162] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

[163] E. L. Thacker and A. Ascherio. Familial aggregation of parkinson's disease: a meta-analysis. *Movement disorders: official journal of the Movement Disorder Society*, 23(8):1174–1183, 2008.

[164] A. Thompson, J. Schäfer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, and C. Hamon. Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by ms/ms. *Analytical chemistry*, 75(8):1895–1904, 2003.

[165] K. Van den Berge, C. Soneson, M. D. Robinson, and L. Clement. stager: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. *Genome biology*, 18(1):151, 2017.

[166] H. D. VanGuilder, K. E. Vrana, and W. M. Freeman. Twenty-five years of quantitative pcr for gene expression analysis. *Biotechniques*, 44(5):619–626, 2008.

[167] D. Velmeshev, L. Schirmer, D. Jung, M. Haeussler, Y. Perez, S. Mayer, A. Bhaduri, N. Goyal, D. H. Rowitch, and A. R. Kriegstein. Single-cell genomics identifies cell type–specific molecular changes in autism. *Science*, 364(6441):685–689, 2019.

[168] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, et al. The sequence of the human genome. *science*, 291(5507):1304–1351, 2001.

[169] D. Virág, B. Dalmadi-Kiss, K. Vékey, L. Drahos, I. Klebovich, I. Antal, and K. Ludányi. Current trends in the analysis of post-translational modifications. *Chromatographia*, 83(1):1–10, 2020.

[170] K. Vitting-Seerup and A. Sandelin. The landscape of isoform switches in human cancers. *Molecular Cancer Research*, 15(9):1206–1220, 2017.

[171] G. M. Wahl, J. L. Meinkoth, and A. R. Kimmel. Northern and southern blots. *Methods in enzymology*, 152, 1987.

[172] C. Wahlestedt. Targeting long non-coding rna to therapeutically upregulate gene expression. *Nature reviews Drug discovery*, 12(6):433–446, 2013.

[173] K. B. Walhovd, H. Johansen-Berg, and R. T. Karadottir. Unraveling the secrets of white matter–bridging the gap between cellular, animal and human imaging studies. *Neuroscience*, 276:2–13, 2014.

[174] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, C. Mayr, S. F. Kingsmore, G. P. Schroth, and C. B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[175] Z. Wang, M. Gerstein, and M. Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, 10(1):57–63, 2009.

[176] C. Ward. Research diagnostic criteria for parkinson's disease. *Advance in Neurology*, 53:245–249, 1990.

[177] M. Weber. The central dogma as a thesis of causal specificity. *History and philosophy of the life sciences*, pages 595–609, 2006.

[178] Y.-N. Wei, H.-Y. Hu, G.-C. Xie, N. Fu, Z.-B. Ning, R. Zeng, and P. Khaitovich. Transcript and protein expression decoupling reveals rna binding proteins and mirnas as potential modulators of human aging. *Genome biology*, 16(1):1–15, 2015.

[179] D. Weisman, M. Cho, C. Taylor, A. Adame, L. Thal, and L. Hansen. In dementia with lewy bodies, braak stage determines phenotype, not lewy body distribution. *Neurology*, 69(4):356–359, 2007.

[180] K. Wirdefeldt, M. Gatz, C. A. Reynolds, C. A. Prescott, and N. L. Pedersen. Heritability of parkinson disease in swedish twins: a longitudinal study. *Neurobiology of aging*, 32(10):1923–e1, 2011.

[181] T. Yasuda and H. Mochizuki. The regulatory role of α-synuclein and parkin in neuronal cell apoptosis; possible implications for the pathogenesis of parkinson's disease. *Apoptosis*, 15(11):1312–1321, 2010.

[182] J. Zaccai, C. Brayne, I. McKeith, F. Matthews, P. Ince, et al. Patterns and stages of α-synucleinopathy: relevance in a population-based cohort. *Neurology*, 70(13):1042–1048, 2008.

[183] N. H. Zawia, D. K. Lahiri, and F. Cardozo-Pelaez. Epigenetics, oxidative stress, and alzheimer disease. *Free radical biology and medicine*, 46(9):1241–1249, 2009.

[184] Y. Zhang, M. James, F. A. Middleton, and R. L. Davis. Transcriptional analysis of multiple brain regions in parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 137(1):5–16, 2005.

[185] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou. Comparison of rna-seq by poly (a) capture, ribosomal rna depletion, and dna microarray for expression profiling. *BMC genomics*, 15(1):1–11, 2014.

[186] Q. Zheng, T. Huang, L. Zhang, Y. Zhou, H. Luo, H. Xu, and X. Wang. Dysregulation of ubiquitin-proteasome system in neurodegenerative diseases. *Frontiers in aging neuroscience*, 8:303, 2016.

[187] L. Zondler, M. Kostka, P. Garidel, U. Heinzelmann, B. Hengerer, B. Mayer, J. H. Weishaupt, F. Gillardon, and K. M. Danzer. Proteasome impairment by α-synuclein. *PloS one*, 12(9):e0184040, 2017.

Chapter 7

Scientific articles

Paper I

Common gene expression signatures in Parkinson's disease are driven by changes in cell composition

Gonzalo S. Nido, Fiona Dick, Lilah Toker, Kjell Petersen, Guido Werner Alves, Ole-Bjørn Tysnes, Inge Jonassen, Kristoffer Haugarvoll & Charalampos Tzoulis
Acta Neuropathologica Communications, **8/55** (2020)

Nido et al. Acta Neuropathologica Communications (2020) 8:55
https://doi.org/10.1186/s40478-020-00932-7

Acta Neuropathologica Communications

RESEARCH Open Access

Common gene expression signatures in Parkinson's disease are driven by changes in cell composition

Gonzalo S. Nido[1,2], Fiona Dick[1,2], Lilah Toker[1,2], Kjell Petersen[1,3], Guido Alves[4,5], Ole-Bjørn Tysnes[1,2], Inge Jonassen[1,3], Kristoffer Haugarvoll[1,2] and Charalampos Tzoulis[1,2*]

Abstract

The etiology of Parkinson's disease is largely unknown. Genome-wide transcriptomic studies in bulk brain tissue have identified several molecular signatures associated with the disease. While these studies have the potential to shed light into the pathogenesis of Parkinson's disease, they are also limited by two major confounders: RNA post-mortem degradation and heterogeneous cell type composition of bulk tissue samples. We performed RNA sequencing following ribosomal RNA depletion in the prefrontal cortex of 49 individuals from two independent case-control cohorts. Using cell type specific markers, we estimated the cell type composition for each sample and included this in our analysis models to compensate for the variation in cell type proportions. Ribosomal RNA depletion followed by capture by random primers resulted in substantially more even transcript coverage, compared to poly(A) capture, in post-mortem tissue. Moreover, we show that cell type composition is a major confounder of differential gene expression analysis in the Parkinson's disease brain. Accounting for cell type proportions attenuated numerous transcriptomic signatures that have been previously associated with Parkinson's disease, including vesicle trafficking, synaptic transmission, immune and mitochondrial function. Conversely, pathways related to endoplasmic reticulum, lipid oxidation and unfolded protein response were strengthened and surface as the top differential gene expression signatures in the Parkinson's disease prefrontal cortex. Our results indicate that differential gene expression signatures in Parkinson's disease bulk brain tissue are significantly confounded by underlying differences in cell type composition. Modeling cell type heterogeneity is crucial in order to unveil transcriptomic signatures that represent regulatory changes in the Parkinson's disease brain and are, therefore, more likely to be associated with underlying disease mechanisms.

Keywords: RNA sequencing, Neurodegeneration, Parkinsonism, Mitochondria, Gene expression

* Correspondence: charalampos.tzoulis@nevro.uib.no; charalampos.tzoulis@helse-bergen.no
[1]Neuro-SysMed Center of Excellence for Clinical Research in Neurological Diseases, Department of Neurology, Haukeland University Hospital, 5021 Bergen, Norway
[2]Department of Clinical Medicine, University of Bergen, Pb 7804, 5020 Bergen, Norway
Full list of author information is available at the end of the article

Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder, affecting ~ 1.8% of the population above 65 years [45]. PD is a complex disorder caused by a combination of genetic and environmental factors, but the molecular mechanisms underlying its etiology remain largely unaccounted for. Genome-wide transcriptomic studies can identify expression signatures associated with PD. While not able to establish causality, these studies hold the potential to highlight important biological mechanisms, some of which may be exploited as targets for therapeutic modulation.

A recent systematic review identified 33 original genome-wide transcriptomic studies in the PD brain, of which 5 were performed on laser microdissected neurons from the *substantia nigra pars compacta* (SNc) and the remaining in bulk tissue from various brain regions [8]. These studies show surprisingly low replicability at the level of individual genes, however, and only partial concordance for pathways. The most consistent alterations have been found in pathways related to energy metabolism/mitochondrial function and protein degradation, followed by synaptic transmission, vesicle trafficking, lysosome/autophagy and neuroinflammation [8]. While these processes commonly show differential expression signatures in PD, it remains unknown whether this is because they truly reflect the biology of PD or due to systematic bias and confounding factors. Two major sources of bias for transcriptomic studies in the human brain are the post-mortem degradation of RNA and the highly heterogeneous cell type composition of bulk tissue samples.

RNA degradation of variable extent occurs in post-mortem tissue. To further complicate the picture, it has been shown that different cell types exhibit different degrees of susceptibility to RNA degradation [32], potentially confounding differences in cellular composition with differences in RNA quality. Access to high-quality brain tissue is generally limited, and thus an optimal choice of experimental platforms becomes paramount to maximize sensitivity. While RNA microarrays are being gradually superseded by RNA-seq technology, only 3 out of the 33 studies identified by an up-to-date review [8] used RNA-seq, and all of them employed poly(A) capture, a widely used protocol (in both RNA-seq and microarray analyses) to restrict the analysis to mature mRNA [20, 30, 46]. However, this library preparation method only picks up RNA fragments with a poly-A tail, introducing substantial bias in low quality RNA samples [1, 25, 47, 56]. A well-established approach to mitigate this limitation is whole RNA-seq following active ribosomal RNA (rRNA) depletion and capture by random primers, such as the Illumina Ribo-Zero technique [31]. To our knowledge there are no genome-wide transcriptomic studies on PD brain employing active rRNA depletion methods to date.

Systematic differences in sample cell composition represent another important confounding factor. These typically originate from two sources: biological differences (e.g. secondary to neurodegeneration) and technical variation in sample dissection and preparation. Brain areas affected by neurodegeneration are characterized by neuronal loss and gliosis, resulting in a systematically increased glia-to-neurons ratio in patients. This confounder is strongest in areas with severe changes, such as the SNc, but is also present to a variable degree in less affected areas, such as the neocortex. In addition, technical sources of variation due to sampling may affect any brain region and cause an uneven distribution of gray and white matter, resulting in a variable fraction of oligodendrocytes. Thus, transcriptional signatures associated with PD in bulk brain tissue may reflect changes in cellular composition rather than disease-specific transcriptional modulation. This observation has already been put forward using neurodegenerative mouse models and re-analysis of human brain transcriptomic data [50]. Heterogeneous cell composition is, hence, a major confounder that needs to be considered and appropriately addressed in transcriptomic studies in bulk brain samples.

We report the first genome-wide transcriptomic study in the PD brain employing RNA-seq following rRNA depletion and random primer capture. We show that this approach is able to substantially mitigate the bias of post-mortem degradation, resulting in substantially better transcript coverage compared to poly(A) capture. Moreover, by estimating the relative cell type proportion in our samples, we confirm that cellular composition is a major source of variation in bulk tissue data, confounding the differential gene expression profile even in the less affected prefrontal cortex. By incorporating the estimated cell type proportions into our analysis models, we were able to unveil transcriptomic signatures which are more likely to be associated with the underlying disease mechanisms.

Material and methods

Subject cohorts

All experiments were conducted in fresh-frozen prefrontal cortex (Brodmann area 9) from a total of 49 individuals from two independent cohorts. The first cohort ($n = 29$) comprised individuals with idiopathic PD ($n = 18$) from the Park-West study (PW), a prospective population-based cohort which has been described in detail [2] and neurologically healthy controls (Ctrl, $n = 11$) from our brain bank for aging and neurodegeneration. Whole-exome sequencing had been performed on all patients [24] and known/predicted pathogenic mutations in genes implicated in Mendelian PD and other

monogenic neurological disorders had been excluded. None of the study participants had clinical signs of mitochondrial disease or use of medication known to influence mitochondrial function (Additional file 1). Controls had no known neurological disease and were matched for age and gender. The second cohort comprised samples from 21 individuals from the Netherlands Brain Bank (NBB) including idiopathic PD ($n = 10$) and demographically matched neurologically healthy controls ($n = 11$). Individuals with PD fulfilled the National Institute of Neurological Disorders and Stroke [26] and the UK Parkinson's disease Society Brain Bank [54] diagnostic criteria for the disease at their final visit. Ethical permission for these studies was obtained from our regional ethics committee (REK 2017/2082, 2010/1700, 131.04).

To investigate the effect of the rRNA depletion and random primer capture protocol compared to the prevailing poly(A) method, we re-analyzed an RNA-seq dataset from a previous publication which employed a poly(A) tail selection kit on post-mortem tissue of the same brain area and same disease (PA cohort, $n = 29$ PD samples, $n = 44$ neurologically healthy controls, all males; GEO: GSE68719) [20]. Informed consent was available from all individuals.

Tissue collection and neuropathology

Brains were collected at autopsy and split sagittaly along the *corpus callosum*. One hemisphere was fixed whole in formaldehyde and the other coronally sectioned and snap-frozen in liquid nitrogen. All samples were collected using a standard technique and fixation time of ~ 2 weeks. There was no significant difference in post-mortem interval (PMI) (independent t-test, PW cohort $p = 0.16$; NBB cohort $p = 0.92$), age (independent t-test, PW cohort $p = 0.18$; NBB cohort $p = 0.074$) or gender (independent t-test, PW cohort $p = 0.94$; NBB cohort $p = 0.53$) between PD subjects and controls. Subject demographics and tissue availability are provided in Additional file 1. Routine neuropathological examination including immunohistochemistry for α-synuclein, tau and beta amyloid was performed on all brains. All cases showed neuropathological changes consistent with PD including degeneration of the dopaminergic neurons of the SNc in the presence of Lewy pathology. Controls had no pathological evidence of neurodegeneration.

RNA sequencing

Total RNA was extracted from prefrontal cortex tissue homogenate for all samples using RNeasy plus mini kit (Qiagen) with on-column DNase treatment according to manufacturer's protocol. Final elution was made in 65 μl of dH2O. The concentration and integrity of the total RNA was estimated by Ribogreen assay (Thermo Fisher Scientific), and Fragment Analyzer (Advanced Analytical), respectively and 500 ng of total RNA was used for downstream RNA-seq applications. First, rRNA was removed using Ribo-Zero™ Gold (Epidemiology) kit (Illumina, San Diego, CA) using manufacturer's recommended protocol. Immediately after the rRNA removal the RNA was fragmented and primed for the first strand synthesis using the NEBNext First Strand synthesis module (New England BioLabs Inc., Ipswich, MA). Directional second strand synthesis was performed using NEBNExt Ultra Directional second strand synthesis kit. Following this the samples were taken into standard library preparation protocol using NEBNext® DNA Library Prep Master Mix Set for Illumina® with slight modifications. Briefly, end-repair was done followed by poly(A) addition and custom adapter ligation. Post-ligated materials were individually barcoded with unique in-house Genomic Services Lab (GSL) primers and amplified through 12 cycles of PCR. Library quantity was assessed by Picogreen Assay (Thermo Fisher Scientific), and the library quality was estimated by utilizing a DNA High Sense chip on a Caliper Gx (Perkin Elmer). Accurate quantification of the final libraries for sequencing applications was determined using the qPCR-based KAPA Biosystems Library Quantification kit (Kapa Biosystems, Inc.). Each library was diluted to a final concentration of 12.5 nM and pooled equimolar prior to clustering. One hundred twenty-five bp Paired-End (PE) sequencing was performed on an Illumina HiSeq2500 sequencer (Illumina, Inc.). RNA quality, as measured by the RNA integrity number (RIN), varied across samples (mean = 5.3, range = 3.0–7.2 for PW; mean = 6.8, range = 3.2–9.1 for NBB), although the difference between conditions did not reach statistical significance in any of the cohorts (t-test $P = 0.72$ and 0.90 for PW and NBB cohorts, respectively).

Data quality control

FASTQ files were trimmed using Trimmomatic version 0.36 [7] to remove potential Illumina adapters and low quality bases with the following parameters: ILLUMINACLIP:truseq.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15. FASTQ files were assessed using fastQC version 0.11.5 [3] prior and following trimming. For an in-depth quality assessment, we mapped the trimmed reads using HISAT2 version 2.1.0 [34] against the hg19 human reference genome (using --rna-strandness RF option) preserving lane-specific information. To discard potential lane-specific sequencing batch effects we inspected the output of the CollectRnaSeqMetrics tool of Picard Tools version 2.6 [11]. Mapping efficiency and proportion of reads mapping to rRNA, intronic, intergenic and coding regions were obtained from the output of the CollectRnaSeqMetrics (Additional file 2: Figure S1 and S2).

For the poly(A) capture dataset [20], raw FASTQ files were obtained from the Gene Expression Omnibus (GEO:GSE68719) and analyzed exactly as described for our cohorts (with the exception of --rna-strandness in HISAT2, which was turned off to take into account that the cDNA library of this cohort was unstranded).

RNA expression quantification and filtering

We used Salmon version 0.9.1 [43] to quantify the abundance at the transcript level with the fragment-level GC bias correction option (–-gcBias) and the appropriate option for the library type (–l ISR) against the Ensembl release 75 transcriptome. Transcript-level quantification was collapsed onto gene-level quantification using the tximport R package version 1.8.0 [49] according to the gene definitions provided by the same Ensembl release. We filtered out genes in non-canonical chromosomes and scaffolds, and transcripts encoded by the mitochondrial genome. To further reduce the potential for artifacts we filtered out transcripts with unusually high expression by removing transcripts that gathered more than 1% of the reads on more than half of the samples, which resulted in the removal of 3 and 4 transcripts from the PW and NBB cohorts, respectively. Additionally, low-expressed (i.e. genes whose expression was below the median expression in at least 20% of the samples) were filtered out from downstream analyses. Samples were then marked as outliers if their median correlation in gene expression (log counts per million) with the other samples was below $Q_1-1.5*IQR$ or above $Q_3 + 1.5*IQR$ (*Tukey's fences*; Q_1: first quartile, Q_3: third quartile, IQR: inter-quartile range). As a result, 3 samples were marked as outliers in the PW cohort and 3 in the NBB cohort, and were not included in downstream analyses (resulting sample sizes: $N_{PW} = 26$, $N_{NBB} = 18$, Additional file 2: Figure S3).

Estimation of marker gene profiles

It has been previously shown that cell type-specific transcriptional signature patterns derived from bulk tissue samples (marker gene profiles, MGPs), can be used as surrogates for relative cell type abundance across samples [37]. MGPs for each cell type are calculated individually, by summarizing the concordant change in their respective marker genes via the first principal component of their expression (i.e. log-transformed counts per million (CPMs)). For the purpose of our study, we calculated MGPs for the main cortical cell types (neurons, oligodendroglia, microglia, endothelial cells, and astrocytes). Cortical cell type markers were obtained from the NeuroExpresso database [37], a comprehensive database compiled using mouse brain cell type expression datasets, and human orthologs were defined using HomoloGene [38]. To reduce the impact of outlier samples, principal component analysis was repeated 100 times on subsampled data, containing an equal number of subjects per group, and removing markers with opposite sign of the main trend. The median score for each sample was used as MGP for the downstream analyses. MGPs obtained with Neuroexpresso-based markers were highly correlated with MGPs calculated using two independent sets of markers from human brain single-cell transcriptomic studies [33, 53] (Additional file 2: Figures S4–8, Additional file 2: Table S1). To assess potential variations associated with the disease across the neuronal markers, we examined the overlap between the markers and the differentially expressed genes in four publicly available datasets of laser microdissected neurons from PD brain (SNc dopaminergic neurons [13, 22, 48] and posterior cingulate cortex pyramidal neurons [51]). We found minimal overlap (3/78 genes) between our neuronal markers and genes differentially expressed in PD dopaminergic neurons. Moreover, none of the markers were differentially expressed in PD cortical neurons [51] (Additional File 2: Figure S9). The vast majority of the cell type markers used for the calculation of MGPs changed in the same direction across our samples (Additional File 2: Figure S9), indicating that MGPs truly represent changes in global cell type-specific transcription profiles, rather than being driven by changes in specific genes.

To unravel potential complex interactions between MGPs and other experimental covariates, including disease status, we calculated the pairwise correlation between all the variables and also their association with the main axes of variation of gene expression. To assist us in choosing an optimal set of MGPs to include as covariates, we quantified the group differences in the cellular proportions between PD and controls using linear models adjusting for the known experimental covariates (i.e. RIN, PMI, sex, age, and sequencing batch). Significant association with disease status was found for oligodendrocyte MGP in the PW cohort and for microglia in the NBB cohort. Thus, these were included in the downstream analyses.

Differential gene expression and functional enrichment analyses

We performed differential gene expression analyses using the DESeq2 R package version 1.22.2 [35] with default parameters. Experimental covariates (sex, age, RIN, PMI, and sequencing batch) as well as oligodendrocyte and microglia MGPs were incorporated into the statistical model. Multiple hypothesis testing was performed with the default automatic filtering of DESeq2 followed by false discovery rate (FDR) calculation by the Benjamini-Hochberg procedure. Analyses were carried out independently for the two cohorts. Genes were

scored according to their significance by transforming the p-values to account for direction of change. For each gene, the up-regulated score was calculated as $S_{\mathrm{up}} = \begin{cases} 1-p/2, LFC < 0 \\ p/2, LFC \geq 0 \end{cases}$, and the down-regulated score as $S_{\mathrm{down}} = 1 - S_{up}$, where LFC corresponds to the log fold change and p to the nominal p-value of the gene. Genes were then tested for enrichment using alternatively $\log(S_{\mathrm{up}})$ and $\log(S_{\mathrm{down}})$ scores employing the gene score resampling method implemented in the ermineR package version 1.0.1 [39], an R wrapper package for ermineJ [27] with the complete Gene Ontology (GO) database annotation [5] to obtain lists of up- and down-regulated pathways for each cohort.

In order to characterize the main biological processes affected by the cell type correction, we scored pathways based on the loss of significance caused by the addition of cellular estimates to the gene expression model. We quantified the difference in the level of significance in the up- and down-regulated enrichment results for each significant pathway as $\Delta = \log(p_0) - \log(p_{CT})$, where p_{CT} and p_0 are the corrected enrichment p-values for the model with cell types (CT) and without (0), respectively. Only pathways that were significant in either one of the models were analyzed in this manner ($p_0 < 0.05$ or $p_{CT} < 0.05$).

The source code for the analyses is available in the GitLab repository (https://git.app.uib.no/neuromics/cell-composition-rna-pd) under the GPL public license v3.0.

Results

Ribo-zero is superior to poly(A) selection in post-mortem brain

We carried out RNA-seq using rRNA depletion and random primer capture (henceforth referred to as Ribo-Zero) in fresh-frozen prefrontal cortex (Brodmann area 9) from a total of 49 individuals from two independent cohorts: the Norwegian ParkWest study (PW, $n = 29$) [2] and the Netherlands Brain Bank (NBB, $n = 21$). Comparison of our data to a published poly(A) capture dataset of similar characteristics [20] (PA cohort) revealed important differences of mapping coverage. Mapping efficiency was slightly higher in the poly(A) dataset (PA: median = 0.976, range = 0.971–0.980) compared to the Ribo-Zero datasets (PW: median = 0.952, range = 0.940–0.962; NBB: median = 0.959, range = 0.947–0.965). The counts per million (CPM) of rRNA regions, as defined by Ensembl release 75, was very low in all samples (PW: median = 3099, range = 1047-7071; NBB: median = 1583, range = 1129-5024) and, as expected, significantly lower in the Ribo-Zero cohorts compared to the poly(A) dataset (PA: median = 40,058, range = 10,701-95,183) (Additional file 2: Figure S1).

In both datasets, the RIN was positively correlated with mapping efficiency to mRNA regions, but not to intergenic and/or intronic regions (Additional file 2: Figure S2). Despite having higher mean RIN values, the PA cohort showed a marked unevenness of transcript body coverage compared to the Ribo-Zero cohorts (Fig. 1a). The median coefficient of variation in coverage was significantly lower in the Ribo-Zero cohorts and the 5′- and 3′-ends of the transcripts showed substantially better coverage compared to the PA cohort (Fig. 1b). Moreover, in the Ribo-Zero datasets both the 3′- and 5′-end coverage loss showed a significant inverse correlation with the RIN values. In contrast, RIN showed no correlation with the 5′-bias and a positive correlation with the 3′-bias in the PA dataset (Fig. 1c). Thus, Ribo-Zero results in substantially better and more even coverage of the transcriptome in post-mortem brain tissue, providing a better alternative to poly(A) capture and minimizing the prospect of transcript quantification biases downstream.

Cell composition is a major confounder of gene expression in bulk brain samples

The observed gene expression profiles in bulk brain tissue can be dramatically influenced by differences in cellular composition. Such differences can be a result of variation in gray/white matter ratios introduced during tissue extraction, inter-subject variability or represent disease related alterations [14, 37, 52]. To study the contribution of various technical and biological sources of variation in our dataset we first estimated marker gene profiles (MGPs) for the major classes of cortical cell types (astrocytes, microglia, oligodendrocytes, endothelial cells and neurons) in our samples by summarizing the expression of the cell type-specific marker genes as previously described [37, 52]. Next, we examined the Pearson's correlation between potential sources of biological variation in our data, including technical and demographic factors (RIN, PMI, sex, age, and disease status) and MGPs. MGPs for neuronal cell types were significantly anticorrelated with the other main cortical cell types in both cohorts ($p < 0.05$, Fig. 2a). In agreement with previous studies [6, 32], MGPs were also correlated with RNA quality. In both cohorts RIN was significantly correlated with neuronal (positive correlation) and astrocyte (negative correlation) MGPs. Significant negative correlation of RIN with microglia MGPs was observed in the NBB cohort (Fig. 2a). Most concerning was the detection of a significant association between the oligodendrocyte MGP and the disease status in the PW cohort (Fig. 2a). The main axis of variation in gene expression (which explained 44 and 45% of the total variance in PW and NBB, respectively) was significantly correlated with RNA quality and cellular composition in both cohorts (Fig. 2b), singling out RNA quality and

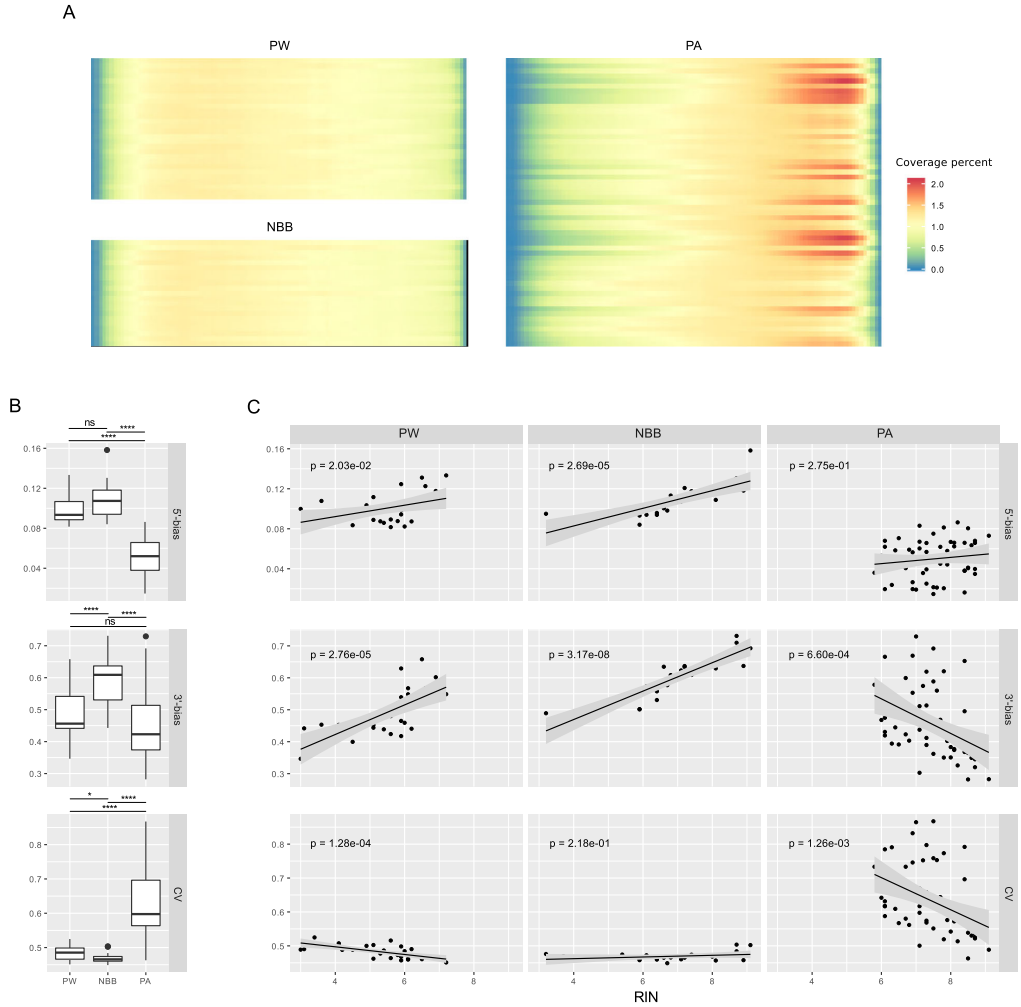


Fig. 1 Transcript coverage profiles of Ribo-Zero datasets compared to poly(A). **a** Heatmaps of transcript coverage in our two cohorts (PW, NBB) and a poly(A) dataset (PA). The y-axis shows samples sorted by RIN (top: lowest RIN; bottom highest RIN). The x-axis represents the transcript body percentiles (5′ to 3′). The shading for a given row represents the sample-normalized coverage averaged across all transcripts. **b** Boxplots for different coverage quality metrics: median 5′-bias, median 3′-bias and median coefficient of variation (CV) for each cohort. The bias metric is calculated by Picard tools on the 1000 most highly expressed transcripts and corresponds to the mean coverage of the 3′(or 5′)-most 100 bases divided by the mean coverage of the whole transcript. Values closer to 1 indicate absence of bias, while values departing from 1 indicate a coverage bias (asterisks indicate significance at (*) $p > 0.05$, (**) $p \leq 0.01$, (***) $p \leq 0.001$, (****) $p \leq 0.0001$, Wilcoxon test). The same metrics are expanded in (**c**), with sample scatterplots showing RIN values against the coverage quality metrics. Linear regression trends are indicated with black lines. *P*-values for the F-statistic of the linear model are also shown in the panels. Panels are organized in columns (cohorts) and quality metrics (rows). CV = coefficient of variation; PW = ParkWest cohort; NBB = Netherlands Brain Bank cohort; PA = poly(A) cohort

cellular composition as the main drivers of transcriptional change in bulk brain tissue.

We next looked for differences in cellular proportions between PD and controls adjusting for the known experimental covariates. In the NBB cohort, PD subjects exhibited a significant increase in the microglia MGP ($p = 0.015$, Wilcoxon test), while a significant increase in the oligodendrocyte MGP ($p = 5.5 \times 10^{-3}$, Wilcoxon test)

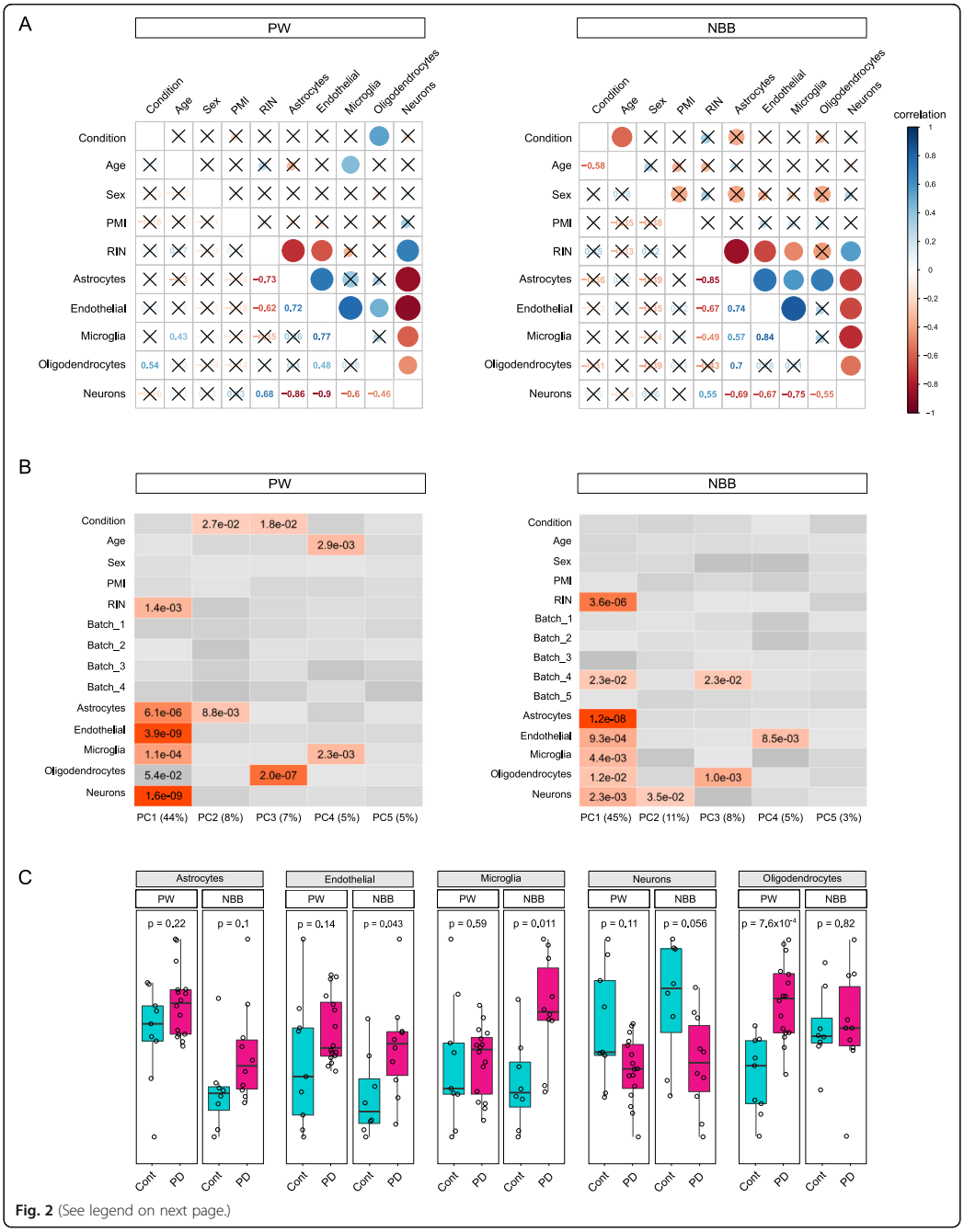


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Analysis of sample covariates. **a** Pearson correlation coefficients for each pair of variables are shown in correlograms. Sizes of the circles in the upper triangular of the correlograms are proportional to the Pearson correlation coefficient, with color indicating positive (blue) or negative (red) coefficients. The precise values for the Pearson coefficients are indicated in the lower triangular. Non-significant pairwise correlations ($p \geq 0.05$) are represented with a cross. **b** Heatmaps showing the association between the sample variables with the first 5 principal components of the gene expression. Only significant *p*-values ($p < 0.05$) are shown (linear regression F-test). **c** Cell type estimates based on MGPs for the main cortical cell types controlling for all the experimental variables except disease status (i.e. sex, age, PMI, RIN, and sequencing batch). P-values calculated with Wilcoxon tests. PW = ParkWest cohort; NBB = Netherlands Brain Bank cohort

was observed in PD subjects from the PW cohort. In both cohorts, these changes were accompanied by a non-significant decrease in neuronal MGPs (Fig. 2c).

MGPs of different cell types are not entirely independent from each other, since changes in one cell type can be accompanied by changes in other cell types. Thus, to ensure that neuronal, endothelial, and astrocyte MGPs do not differ between the groups, we re-estimated group differences in these MGPs while adjusting for the oligodendrocyte and microglia MGPs. This analysis showed no significant differences between the groups (Additional file 2: Figure S10). Therefore, only MGPs of oligodendrocytes and microglia were included in the statistical model of differential expression.

Differential gene expression

Differential gene expression analysis of a total of ~ 31,000 pre-filtered genes was carried out using experimental covariates (sex, age, PMI, RIN, and sequencing batch) with or without oligodendrocyte and microglia MGPs. In the PW cohort, 595 genes were defined as differentially expressed (FDR < 0.05) without adjusting for cell type composition. Inclusion of oligodendrocyte and microglia MGPs in the model decreased the number of differentially expressed genes to a total of 220. In total, 74 genes remained significant both with and without adjustment for cell type composition. No genes with FDR < 0.05 were identified in the NBB cohort, irrespective of adjustment for cell type composition. A list with the nominally significant genes overlapping between the two cohorts is provided in Additional file 3. Comprehensive results of differential expression analysis are available in Additional file 4.

Functional enrichment

Functional enrichment analysis of the differential gene expression results without MGP adjustment indicated 476 significantly enriched (FDR < 0.05) pathways in PW (107 up-regulated and 369 down-regulated) and 992 in NBB (421 up-regulated and 571 down-regulated). MGP adjustment reduced the number of significant pathways to 89 in PW (35 up-regulated and 54 down-regulated) and 248 in NBB (115 up-regulated and 133 down-regulated). Of these, 34 pathways replicated across the two cohorts. Concordant pathways comprised protein folding, ER-related processes and lipid oxidation (Fig. 3). The complete results are provided in Additional file 5.

As expected, scoring each pathway according to the change in *p*-value when accounting for cellularity, revealed a marked downplay of the relevant cell type-specific functions (Table 1). In the PW cohort, which was characterized by a skewed oligodendrocytes/neurons proportion, the function with the largest attenuation (i.e. increase in p-value) was seen for up-regulation of myelination and other oligodendrocyte related functions and for down-regulation of neuronal pathways. For NBB, accounting for cell-composition resulted in attenuation of immunity and neuronal pathways, consistent with the unbalanced microglial/neuronal proportions seen in that cohort (Table 1). Strikingly, pathways linked to mitochondrial respiration, including respiratory complex I, were among the down-regulated processes that lost statistical significance when controlling for cellularity. The attenuation of the mitochondrial signal was observed in both cohorts. Conversely, up-regulation of protein folding-related pathways gained significance in both cohorts (Table 2 and Fig. 3). Complete results are provided in Additional file 6.

Discussion

We present the first genome-wide transcriptomic study in the PD brain employing whole RNA-seq after rRNA depletion and random primer capture (Ribo-Zero). Our findings show that PD-associated differential gene expression signatures in bulk brain tissue are influenced to a great extent by the underlying differences in cell type composition of the samples. Modeling cell type heterogeneity allowed us to highlight transcriptional signatures that are likely to represent aberrant gene expression within the cells of the PD brain, rather than changes in cell composition.

Our results suggest that the Ribo-Zero approach is superior to the more commonly used poly(A) method and allows for a more accurate mapping and quantification of the transcriptome in post-mortem brain tissue. The Ribo-Zero method provides substantially higher evenness of coverage and effectively mitigates the 3′- and 5′-end coverage bias associated with poly(A) capture. Ultimately, the unevenness of coverage will influence transcript quantification, affecting the sensitivity of the

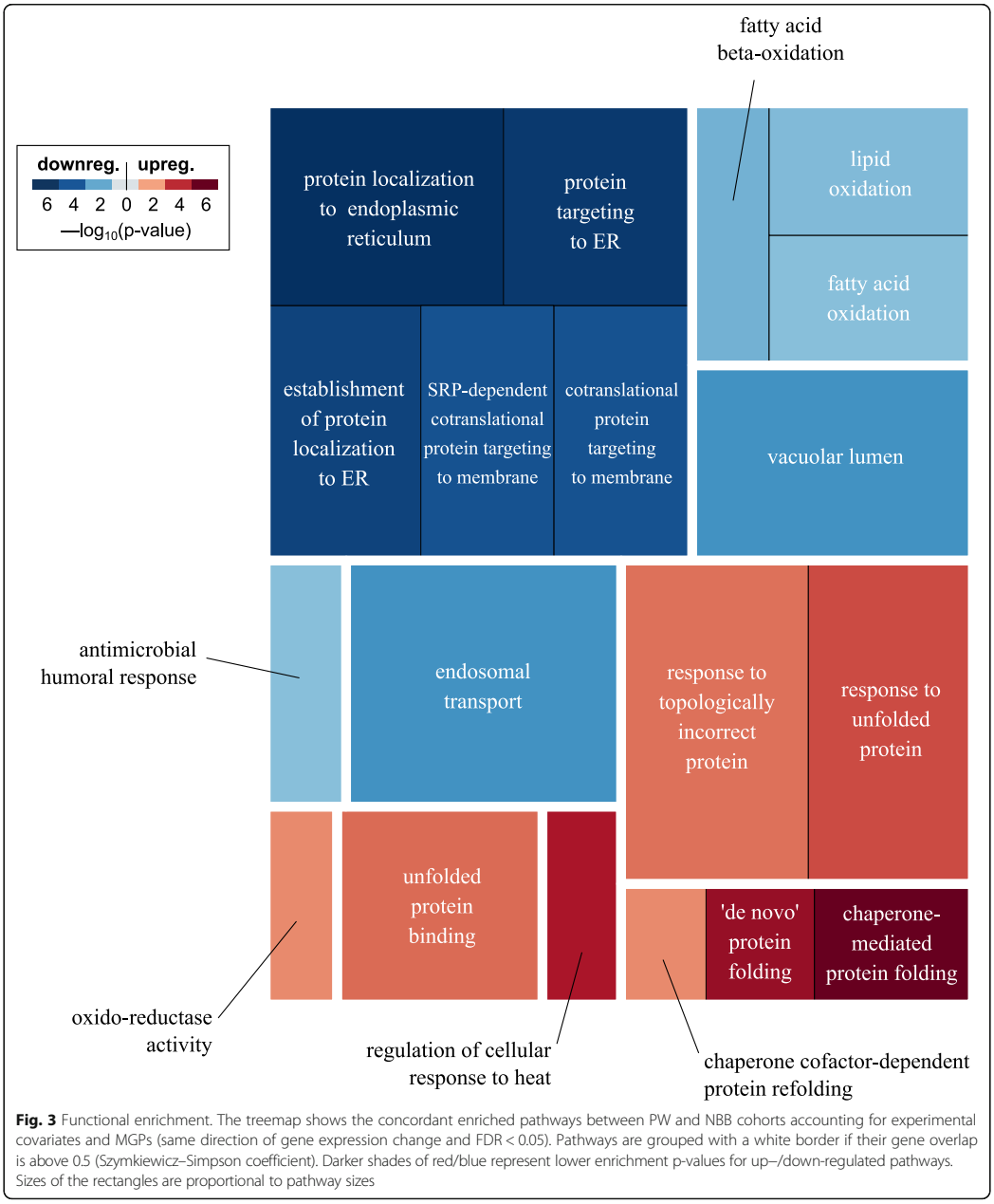


Fig. 3 Functional enrichment. The treemap shows the concordant enriched pathways between PW and NBB cohorts accounting for experimental covariates and MGPs (same direction of gene expression change and FDR < 0.05). Pathways are grouped with a white border if their gene overlap is above 0.5 (Szymkiewicz–Simpson coefficient). Darker shades of red/blue represent lower enrichment p-values for up–/down-regulated pathways. Sizes of the rectangles are proportional to pathway sizes

differential expression estimates. While these observations are in agreement with previous comprehensive reports [1, 25, 47, 56], we cannot rule out the contribution of experimental variables specific to each cohort, in addition to the RNA sequencing methodology. Furthermore, while the Ribo-Zero protocol shows advantages

Table 1 Loss of significance in enriched pathways

PW

Up-regulated		Down-regulated	
Pathway	***Delta***	***Pathway***	***Delta***
myelination	−8.85	regulation of synaptic vesicle exocytosis	−9.63
ensheathment of neurons	−8.67	intrinsic component of synaptic membrane	−9.60
axon ensheathment	−8.67	regulation of synaptic vesicle cycle	−9.58
detection of chemical stimulus involved in sensory perception of bitter taste	−7.95	positive regulation of synaptic transmission	−9.48
oligodendrocyte differentiation	−6.94	Schaffer collateral - CA1 synapse	−9.46
oligodendrocyte development	−6.58	regulation of synaptic plasticity	−9.44
apical junction complex	−5.57	regulation of neurotransmitter secretion	−9.41
glial cell development	−5.54	presynaptic membrane	−9.27
glial cell differentiation	−5.52	regulation of synaptic vesicle transport	−9.19
tight junction	−4.16	protein transport within lipid bilayer	−9.11

NBB

Up-regulated		Down-regulated	
Pathway	***Delta***	***Pathway***	***Delta***
activation of innate immune response	−8.54	ribonucleoside monophosphate metabolic process	−9.09
regulation of leukocyte proliferation	−8.11	purine nucleoside triphosphate metabolic process	−9.03
regulation of lymphocyte proliferation	−8.01	mitochondrial membrane part	−8.99
regulation of mononuclear cell proliferation	−7.79	ATP metabolic process	−8.92
innate immune response-activating signal transduction	−7.43	regulation of synaptic vesicle exocytosis	−8.87
regulation of adaptive immune response	−7.16	inner mitochondrial membrane protein complex	−8.82
response to interferon-gamma	−7.15	purine ribonucleoside triphosphate metabolic process	−8.77
adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	−7.07	respiratory chain	−8.70
blood microparticle	−6.75	regulation of synaptic vesicle transport	−8.68
regulation of T cell proliferation	−6.73	cellular respiration	−8.43

Tables representing the top 10 pathways with the lowest delta for up- and down-regulated pathways for PW and NBB cohorts. The delta value represents the change in the enrichment $-\log_{10}$ (*p*-value) between the results with and without MGP adjustment (negative values of delta imply a loss of significance when accounting for cellularity). Complete results are provided in Additional file 6

compared to the poly(A) method, it is certainly not sufficient to fully mitigate the impact of RNA degradation on transcript quantification.

Our study supports the notion that cell composition can be a major confounder in bulk brain tissue transcriptomics. We estimated the relative cell type abundance across our samples by calculating MGPs for the main cortical cell types. While MGPs do not provide a direct measure of cell counts, they are a validated and robust surrogate for cell type composition [37, 52]. Moreover, we show that MGPs are (1) highly consistent across three different single cell-based marker sets, (2) highly robust to marker gene outliers, and (3) not susceptible to PD-associated changes in gene expression. Taken together, these results indicate that MGPs reliably represent the general behavior of cell type-specific transcriptional signature in our data.

Our analyses indicate that the observed expression profiles in both cohorts were driven predominantly by a combination of technical factors associated with RNA quality, and differences in cellular composition between PD and controls. This difference was primarily due to oligodendrocytes in PW and microglia in NBB. Since oligodendrocyte proliferation is not a pathological feature of PD, it is plausible that the difference in oligodendrocyte MGPs in PW was due to technical variation in gray/white matter content introduced during tissue sampling. Microglial infiltration does occur in affected areas of the PD brain [18]. It is noteworthy, however, that increased microglial MGP was only observed in one of the cohorts (NBB), highlighting the biological heterogeneity of PD. Accounting for relative cell proportions reduced the number of differentially expressed genes and attenuated the calculated enrichment of cell type-specific pathways between PD and controls. In the PW cohort, this alleviated a substantial false positive signature of oligodendrocyte genes presumably caused by skewed grey/white matter sampling bias. Similar sampling bias could be responsible for oligodendroglia-specific functions appointed to PD brain in previous transcriptomic studies [46].

Intriguingly, accounting for cellular proportions downplayed several of the transcriptomic signatures that have been previously associated with PD. For instance, the signal from vesicle trafficking- and synaptic transmission-related processes [9, 10, 15, 21, 29, 40] was significantly attenuated in both cohorts, suggesting that the signal was primarily driven by changes in neuronal proportions between PD and controls, rather than modulation of these pathways within neurons. Moreover, we observed an attenuation in the down-regulation of mitochondrial pathways, including the respiratory chain and oxidative phosphorylation, which are among the most consistent transcriptomic signatures in PD [8,

Table 2 Gain of significance in enriched pathways

PW			
Up-regulated		**Down-regulated**	
Pathway	***Delta***	***Pathway***	***Delta***
protein folding	5.77	DNA packaging complex	3.76
'de novo' protein folding	5.73	basement membrane	3.47
unfolded protein binding	5.54	positive regulation of epithelial cell proliferation	2.52
chaperone-mediated protein folding	5.32	negative regulation of gliogenesis	2.49
'de novo' posttranslational protein folding	4.68	fatty acid beta-oxidation	2.30
heat shock protein binding	4.34	nucleosome	2.18
response to unfolded protein	4.10	glomerulus development	2.16
response to topologically incorrect protein	3.53	aorta development	2.06
oxidoreductase activity, acting on paired…	2.74	endothelium development	1.95
NBB			
Up-regulated		**Down-regulated**	
Pathway	***Delta***	***Pathway***	***Delta***
positive regulation of cardiac muscle tissue dev…	1.99	tertiary granule	5.22
regulation of smooth muscle cell differentiation	1.98	ficolin-1-rich granule membrane	5.00
negative regulation of protein serine/threonine kin…	1.98	regulation of myeloid leukocyte mediated immunity	4.55
hormone-mediated signaling pathway	1.95	regulation of leukocyte degranulation	4.34
lung alveolus development	1.71	specific granule	4.22
positive regulation of striated muscle tissue dev…	1.69	ficolin-1-rich granule	4.15
positive regulation of muscle organ development	1.69	tertiary granule membrane	3.57
positive regulation of muscle tissue development	1.62	regulation of mast cell activation	3.52
negative regulation of MAP kinase activity	1.48	vacuolar lumen	3.05
regulation of cardiac muscle cell differentiation	1.48	regulation of mast cell degranulation	3.05

Tables representing the top 10 pathways with the highest delta for up- and down-regulated pathways for both cohorts. The delta value represents the change in the enrichment $-\log_{10}$ (*p*-value) between the results with and without MGP adjustment (positive values imply an increase in *p*-value when accounting for cellularity). Complete results are provided in Additional file 6

9, 19, 20, 29, 42, 55, 57]. The loss of transcriptional signal in these pathways is intriguing, because there is compelling evidence that decreased complex I protein levels occur in PD neurons [23]. Our results suggest that the previously reported transcriptional down-regulation of the respiratory chain is at least partly driven by altered cellular composition (due to decreased number of neurons which highly express these genes) and may therefore not be the sole mechanism by which neuronal complex I deficiency occurs in PD. Indeed, it has been suggested that complex I deficiency in PD may be mediated by proteolytic degradation by the LON-ClpP protease system, rather than transcriptional regulation [44].

Changes in the cell-composition of the affected brain regions occur in all neurodegenerative diseases, including PD, Alzheimer disease, amyotrophic lateral sclerosis (ALS) and Huntington disease. Interestingly, common and overlapping transcriptional signatures have been reported across these neurodegenerative diseases, including mitochondrial, neuronal-specific, and immunity-related pathways [4, 17]. Our findings suggest that these common transcriptional signatures of neurodegeneration may largely represent the common pattern of altered cellularity, involving neuronal loss and glial proliferation, rather than biological processes of causal nature.

Accounting for cell type composition in our samples highlighted processes related to the endoplasmic reticulum, unfolded protein response and lipid/fatty acid oxidation as the top differential gene expression signatures in the PD prefrontal cortex. Unfolded protein response is indeed one of the most consistently reported transcriptomic signatures in PD [8, 9, 20, 28, 41, 55]. Moreover, endoplasmic reticulum stress and aberrant proteostasis have been associated with the accumulation of misfolded proteins, including α-synuclein, in both in vitro studies and animal models of PD [16]. While less is known regarding the role of lipid metabolism in PD, evidence of aberrant fatty acid oxidation has been found by metabolomic studies in serum [12] and urine [36] of patients. Our results corroborate these findings and indicate that aberrant fatty acid metabolism occurs in the PD prefrontal cortex.

Based on our findings, we advocate that modeling cell type heterogeneity is crucial in order to unveil transcriptomic signatures reflecting regulatory changes in the PD brain. It is, however, important noting that modeling of cellular estimates cannot completely mitigate the cell-composition bias in bulk tissue. Moreover, cell type correction complicates the identification of transcriptional changes that are confounded with changes in cellular composition and may thus increase the false negative rate. Single-cell or cell-sorting based methods will be key to overcoming this limitation and deciphering transcriptomic signatures directly associated with underlying disease mechanisms in PD.

Conclusions

Our findings show that differential gene expression signatures derived from bulk brain tissue of PD patients are significantly confounded by underlying differences in cell type composition. Modeling cell type heterogeneity is

crucial in order to unveil transcriptomic signatures that represent regulatory changes in the PD brain and are, therefore, more likely to be associated with underlying disease mechanisms.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s40478-020-00932-7.

Additional file 1:. Cohort demographic and experimental information.

Additional file 2: Figure S1. Read mapping efficiency; **Figure S2.** Read mapping statistics; **Figure S3.** Sample clustering; **Figure S4.** Neuronal MGPs and expression of neuronal markers; **Figure S5.** Oligodendrocyte MGPs and expression of oligodendrocyte markers; **Figure S6.** Microglial MGPs and expression of microglial markers; **Figure S7.** Astrocyte MGPs and expression of astrocyte markers; **Figure S8.** Endothelial MGPs and expression of endothelial markers; **Figure S9.** Neuroexpresso neuronal markers and overlap with single-cell neuronal DEGs; **Figure S10.** Cellular estimates grouped by status; **Table S1.** Correlations between MGPs calculated on different marker sets.

Additional file 3. Significant genes overlapping PW and NBB accounting for cell types.

Additional file 4. Complete results of the differential expression analyses.

Additional file 5. ErmineR pathway enrichment analyses before and after accounting for cell composition.

Additional file 6 Up- and down-regulated pathways ranked by the change in *p*-value resulting from accounting for cell composition.

Additional file 7. Read count matrix.

Abbreviations

ALS: Amyotrophic lateral sclerosis; FDR: Benjamini-Hochberg false discovery rate; GO: Gene Ontology; MGP: Marker gene profiles; NBB: Netherlands Brain Bank; PD: Parkinson's disease; PMI: Post-mortem interval; PW: Park West; RIN: RNA integrity number; rRNA: ribosomal RNA; SNc: *Substantia nigra pars compacta*

Acknowledgments

Not applicable.

Authors' contributions

GSN participated in the study conception, preprocessed the RNA-seq data, designed and performed the RNA-seq computational analyses. FD participated in the RNA-seq computational analyses. KH contributed to data interpretation and critical revision of the manuscript. IJ and KP participated in the conception and design of the study, data interpretation and critical revision of the manuscript. OBT and GA contributed part of the tissue material, supplementary data, and critical revision of the manuscript. CT conceived, designed and directed the study, contributed to data analyses and interpretation and acquired funding for the study. GSN and CT wrote the manuscript with the active input and participation of LT, KH, IJ, and KP All authors have read and approved the manuscript.

Funding

This work is supported by grants from The Research Council of Norway (288164, ES633272) and Bergen Research Foundation (BFS2017REK05).

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its supplementary files. The source code for the analyses is available in the GitLab repository (https://git.app.uib.no/neuromics/cell-composition-rna-pd) under the GPL public license v3.0.

Ethics approval and consent to participate

Informed consent was available from all individuals. Ethical permission for these studies was obtained from our regional ethics committee (REK 2017/2082, 2010/1700, 131.04).

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Neuro-SysMed Center of Excellence for Clinical Research in Neurological Diseases, Department of Neurology, Haukeland University Hospital, 5021 Bergen, Norway. [2]Department of Clinical Medicine, University of Bergen, Pb 7804, 5020 Bergen, Norway. [3]Computational Biology Unit, Department of Informatics, University of Bergen, Pb 7803, 5020 Bergen, Norway. [4]The Norwegian Centre for Movement Disorders and Department of Neurology, Stavanger University Hospital, Pb 8100, 4068 Stavanger, Norway. [5]Department of Mathematics and Natural Sciences, University of Stavanger, 4062 Stavanger, Norway.

Received: 11 March 2020 Accepted: 14 April 2020
Published online: 21 April 2020

References

1. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T, Gnirke A, Pochet N, Regev A, Levin JZ (2013) Comparative analysis of RNA sequencing methods for degraded or low-input samples. Nat Methods 10:623–629. https://doi.org/10.1038/nmeth.2483
2. Alves G, Muller B, Herlofson K, HogenEsch I, Telstad W, Aarsland D, Tysnes O-B, Larsen JP, for the Norwegian ParkWest study group (2009) Incidence of Parkinson's disease in Norway: the Norwegian ParkWest study. J Neurol Neurosurg Psychiatry 80:851–857. https://doi.org/10.1136/jnnp.2008.168211
3. Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc
4. Arneson D, Zhang Y, Yang X, Narayanan M (2018) Shared mechanisms among neurodegenerative diseases: from genetic factors to gene networks. J Genet 97:795–806
5. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29. https://doi.org/10.1038/75556
6. Bauer M (2007) RNA in forensic science. Forensic Sci Int Genet 1:69–74. https://doi.org/10.1016/j.fsigen.2006.11.002
7. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
8. Borrageiro G, Haylett W, Seedat S, Kuivaniemi H, Bardien S (2018) A review of genome-wide transcriptomics studies in Parkinson's disease. Eur J Neurosci 47:1–16. https://doi.org/10.1111/ejn.13760
9. Bossers K, Meerhoff G, Balesar R, van Dongen JW, Kruse CG, Swaab DF, Verhaagen J (2009) Analysis of gene expression in Parkinson's disease: possible involvement of neurotrophic support and axon guidance in dopaminergic cell death. Brain Pathol Zurich Switz 19:91–107. https://doi.org/10.1111/j.1750-3639.2008.00171.x
10. Botta-Orfila T, Tolosa E, Gelpi E, Sànchez-Pla A, Martí M-J, Valldeoriola F, Fernández M, Carmona F, Ezquerra M (2012) Microarray expression analysis in idiopathic and LRRK2-associated Parkinson's disease. Neurobiol Dis 45:462–468. https://doi.org/10.1016/j.nbd.2011.08.033
11. Broad Institute (2018). Picard tools. Available online at:http://broadinstitute.github.io/picard/
12. Burté F, Houghton D, Lowes H, Pyle A, Nesbitt S, Yarnall A, Yu-Wai-Man P, Burn DJ, Santibanez-Koref M, Hudson G (2017) Metabolic profiling of Parkinson's disease and mild cognitive impairment. Mov Disord Off J Mov Disord Soc 32:927–932. https://doi.org/10.1002/mds.26992
13. Cantuti-Castelvetri I, Keller-McGandy C, Bouzou B, Asteris G, Clark TW, Frosch MP, Standaert DG (2007) Effects of gender on nigral gene expression and parkinson disease. Neurobiol Dis 26:606–614. https://doi.org/10.1016/j.nbd.2007.02.009

14. Capurro A, Bodea L-G, Schaefer P, Luthi-Carter R, Perreau VM (2014) Computational deconvolution of genome wide expression data from Parkinson's and Huntington's disease brain tissues using population-specific expression analysis. Front Neurosci 8: 441. https://doi.org/10.3389/fnins.2014.00441
15. Chandrasekaran S, Bonchev D (2013) A network view on Parkinson's disease. Comput Struct Biotechnol J 7:e201304004. https://doi.org/10.5936/csbj.201304004
16. Colla E (2019) Linking the endoplasmic reticulum to Parkinson's disease and alpha-Synucleinopathy. Front Neurosci 13:560. https://doi.org/10.3389/fnins.2019.00560
17. Cooper-Knock J, Kirby J, Ferraiuolo L, Heath PR, Rattray M, Shaw PJ (2012) Gene expression profiling in human neurodegenerative disease. Nat Rev Neurol 8:518–530. https://doi.org/10.1038/nrneurol.2012.156
18. Dickson DW (2012) Parkinson's disease and parkinsonism: neuropathology. Cold Spring Harb Perspect Med 2:a009258–a009258. https://doi.org/10.1101/cshperspect.a009258
19. Duke DC, Moran LB, Kalaitzakis ME, Deprez M, Dexter DT, Pearce RKB, Graeber MB (2006) Transcriptome analysis reveals link between proteasomal and mitochondrial pathways in Parkinson's disease. Neurogenetics 7:139–148. https://doi.org/10.1007/s10048-006-0033-5
20. Dumitriu A, Golji J, Labadorf AT, Gao B, Beach TG, Myers RH, Longo KA, Latourelle JC (2016) Integrative analyses of proteomics and RNA transcriptomics implicate mitochondrial processes, protein folding pathways and GWAS loci in Parkinson disease. BMC Med Genet 9:5. https://doi.org/10.1186/s12920-016-0164-y
21. Edwards YJK, Beecham GW, Scott WK, Khuri S, Bademci G, Tekin D, Martin ER, Jiang Z, Mash DC, ffrench-Mullen J, Pericak-Vance MA, Tsinoremas N, Vance JM (2011) Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. PLoS One 6:e16917. https://doi.org/10.1371/journal.pone.0016917
22. Elstner M, Morris CM, Heim K, Bender A, Mehta D, Jaros E, Klopstock T, Meitinger T, Turnbull DM, Prokisch H (2011) Expression analysis of dopaminergic neurons in Parkinson's disease and aging links transcriptional dysregulation of energy metabolism to cell death. Acta Neuropathol (Berl) 122:75–86. https://doi.org/10.1007/s00401-011-0828-9
23. Flønes IH, Fernandez-Vizarra E, Lykouri M, Brakedal B, Skeie GO, Miletic H, Lilleng PK, Alves G, Tysnes O-B, Haugarvoll K, Dölle C, Zeviani M, Tzoulis C (2018) Neuronal complex I deficiency occurs throughout the Parkinson's disease brain, but is not associated with neurodegeneration or mitochondrial DNA damage. Acta Neuropathol (Berl) 135:409–425. https://doi.org/10.1007/s00401-017-1794-7
24. Gaare JJ, Nido GS, Sztromwasser P, Knappskog PM, Dahl O, Lund-Johansen M, Maple-Grødem J, Alves G, Tysnes O-B, Johansson S, Haugarvoll K, Tzoulis C (2018) Rare genetic variation in mitochondrial pathways influences the risk for Parkinson's disease: mitochondrial pathways in PD. Mov Disord 33: 1591–1600. https://doi.org/10.1002/mds.64
25. Gallego Romero I, Pai AA, Tung J, Gilad Y (2014) RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol 12:42. https://doi.org/10.1186/1741-7007-12-42
26. Gelb DJ, Oliver E, Gilman S (1999) Diagnostic criteria for Parkinson disease. Arch Neurol 56:33–39
27. Gillis J, Mistry M, Pavlidis P (2010) Gene function analysis in complex data sets using ErmineJ. Nat Protoc 5:1148–1159. https://doi.org/10.1038/nprot.2010.78
28. Grünblatt E, Mandel S, Jacob-Hirsch J, Zeligson S, Amariglio N, Rechavi G, Li J, Ravid R, Roggendorf W, Riederer P, Youdim MBH (2004) Gene expression profiling of parkinsonian substantia nigra pars compacta; alterations in ubiquitin-proteasome, heat shock protein, iron and oxidative stress regulated proteins, cell adhesion/cellular matrix and vesicle trafficking genes. J Neural Transm Vienna Austria 111:1543–1573. https://doi.org/10.1007/s00702-004-0212-1
29. Hauser MA, Li Y-J, Xu H, Noureddine MA, Shao YS, Gullans SR, Scherzer CR, Jensen RV, McLaurin AC, Gibson JR, Scott BL, Jewett RM, Stenger JE, Schmechel DE, Hulette CM, Vance JM (2005) Expression profiling of substantia nigra in Parkinson disease, progressive supranuclear palsy, and frontotemporal dementia with parkinsonism. Arch Neurol 62:917–921. https://doi.org/10.1001/archneur.62.6.917
30. Henderson-Smith A, Corneveaux JJ, De Both M, Cuyugan L, Liang WS, Huentelman M, Adler C, Driver-Dunckley E, Beach TG, Dunckley TL (2016) Next-generation profiling to identify the molecular etiology of Parkinson dementia. Neurol Genet 2:e75. https://doi.org/10.1212/NXG.0000000000000075
31. Huang R, Jaritz M, Guenzl P, Vlatkovic I, Sommer A, Tamir IM, Marks H, Klampfl T, Kralovics R, Stunnenberg HG, Barlow DP, Pauler FM (2011) An RNA-Seq strategy to detect the complete coding and non-coding Transcriptome including full-length imprinted macro ncRNAs. PLoS One 6: e27288. https://doi.org/10.1371/journal.pone.0027288
32. Jaffe AE, Tao R, Norris AL, Kealhofer M, Nellore A, Shin JH, Kim D, Jia Y, Hyde TM, Kleinman JE, Straub RE, Leek JT, Weinberger DR (2017) qSVA framework for RNA quality correction in differential expression analysis. Proc Natl Acad Sci U S A 114:7130–7135. https://doi.org/10.1073/pnas.1617384114
33. Kelley KW, Nakao-Inoue H, Molofsky AV, Oldham MC (2018) Variation among intact tissue samples reveals the core transcriptional features of human CNS cell classes. Nat Neurosci 21:1171–1184. https://doi.org/10.1038/s41593-018-0216-z
34. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. https://doi.org/10.1038/nmeth.3317
35. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550. https://doi.org/10.1186/s13059-014-0550-8
36. Luan H, Liu L-F, Meng N, Tang Z, Chua K-K, Chen L-L, Song J-X, Mok VCT, Xie L-X, Li M, Cai Z (2015) LC-MS-based urinary metabolite signatures in idiopathic Parkinson's disease. J Proteome Res 14:467–478. https://doi.org/10.1021/pr500807t
37. Mancarci BO, Toker L, Tripathy SJ, Li B, Rocco B, Sibille E, Pavlidis P (2017, 2017) Cross-laboratory analysis of brain cell type Transcriptomes with applications to interpretation of bulk tissue data. eNeuro 4. https://doi.org/10.1523/ENEURO.0212-17.2017
38. Mancarci O. (2019). Homologene: quick access to homologene and gene annotation updates. R package version 1.4.68. Available online at:https://CRAN.R-project.org/package=homologene
39. Mancarci O. (2019). ermineR: gene set analysis with multifunctionality assessment. R package version 1.0.1. Available online at:https://github.com/PavlidisLab/ermineR
40. Miller RM, Kiser GL, Kaysser-Kranich TM, Lockner RJ, Palaniappan C, Federoff HJ (2006) Robust dysregulation of gene expression in substantia nigra and striatum in Parkinson's disease. Neurobiol Dis 21:305–313. https://doi.org/10.1016/j.nbd.2005.07.010
41. Moran LB, Duke DC, Deprez M, Dexter DT, Pearce RKB, Graeber MB (2006) Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics 7:1–11. https://doi.org/10.1007/s10048-005-0020-2
42. Papapetropoulos S, Ffrench-Mullen J, McCorquodale D, Qin Y, Pablo J, Mash DC (2006) Multiregional gene expression profiling identifies MRPS6 as a possible candidate gene for Parkinson's disease. Gene Expr 13:205–215
43. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14: 417–419. https://doi.org/10.1038/nmeth.4197
44. Pryde KR, Taanman JW, Schapira AH (2016) A LON-ClpP Proteolytic Axis degrades complex I to extinguish ROS production in depolarized mitochondria. Cell Rep 17:2522–2531. https://doi.org/10.1016/j.celrep.2016.11.027
45. de Rijk MC, Launer LJ, Berger K, Breteler MM, Dartigues JF, Baldereschi M, Fratiglioni L, Lobo A, Martinez-Lage J, Trenkwalder C, Hofman A (2000) Prevalence of Parkinson's disease in Europe: a collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. Neurology 54:S21–S23
46. Riley BE, Gardai SJ, Emig-Agius D, Bessarabova M, Ivliev AE, Schüle B, Schüle B, Alexander J, Wallace W, Halliday GM, Langston JW, Braxton S, Yednock T, Shaler T, Johnston JA (2014) Systems-based analyses of brain regions functionally impacted in Parkinson's disease reveals underlying causal mechanisms. PLoS One 9:e102909. https://doi.org/10.1371/journal.pone.0102909
47. Schuierer S, Carbone W, Knehr J, Petitjean V, Fernandez A, Sultan M, Roma G (2017) A comprehensive assessment of RNA-seq protocols for degraded and low-quantity samples. BMC Genomics 18:442. https://doi.org/10.1186/s12864-017-3827-y
48. Simunovic F, Yi M, Wang Y, Stephens R, Sonntag KC (2010) Evidence for gender-specific transcriptional profiles of nigral dopamine neurons in Parkinson disease. PLoS One 5:e8856. https://doi.org/10.1371/journal.pone.0008856

49. Soneson C, Love MI, Robinson MD (2016) Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Res 4:1521. https://doi.org/10.12688/f1000research.7563.2
50. Srinivasan K, Friedman BA, Larson JL, Lauffer BE, Goldstein LD, Appling LL, Borneo J, Poon C, Ho T, Cai F, Steiner P, van der Brug MP, Modrusan Z, Kaminker JS, Hansen DV (2016) Untangling the brain's neuroinflammatory and neurodegenerative transcriptional responses. Nat Commun 7:11295. https://doi.org/10.1038/ncomms11295
51. Stamper C, Siegel A, Liang WS, Pearson JV, Stephan DA, Shill H, Connor D, Caviness JN, Sabbagh M, Beach TG, Adler CH, Dunckley T (2008) Neuronal gene expression correlates of Parkinson's disease with dementia. Mov Disord Off J Mov Disord Soc 23:1588–1595. https://doi.org/10.1002/mds.22184
52. Toker L, Mancarci BO, Tripathy S, Pavlidis P (2018) Transcriptomic evidence for alterations in astrocytes and Parvalbumin interneurons in subjects with bipolar disorder and schizophrenia. Biol Psychiatry 84:787–796. https://doi.org/10.1016/j.biopsych.2018.07.010
53. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, Bhaduri A, Goyal N, Rowitch DH, Kriegstein AR (2019) Single-cell genomics identifies cell type-specific molecular changes in autism. Science 364:685–689. https://doi.org/10.1126/science.aav8130
54. Ward CD, Gibb WR (1990) Research diagnostic criteria for Parkinson's disease. Adv Neurol 53:245–249
55. Zhang Y, James M, Middleton FA, Davis RL (2005) Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am J Med Genet B Neuropsychiatr Genet 137B:5–16. https://doi.org/10.1002/ajmg.b.30195
56. Zhao W, He X, Hoadley KA, Parker JS, Hayes D, Perou CM (2014) Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. BMC Genomics 15:419. https://doi.org/10.1186/1471-2164-15-419
57. Zheng B, Liao Z, Locascio JJ, Lesniak KA, Roderick SS, Watt ML, Eklund AC, Zhang-James Y, Kim PD, Hauser MA, Grünblatt E, Moran LB, Mandel SA, Riederer P, Miller RM, Federoff HJ, Wüllner U, Papapetropoulos S, Youdim MB, Cantuti-Castelvetri I, Young AB, Vance JM, Davis RL, Hedreen JC, Adler CH, Beach TG, Graeber MB, Middleton FA, Rochet J-C, Scherzer CR, Global PD Gene Expression (GPEX) Consortium (2010) PGC-1α, a potential therapeutic target for early intervention in Parkinson's disease. Sci Transl Med 2:52ra73. https://doi.org/10.1126/scitranslmed.3001059

Publisher's Note

Paper II

Differential transcript usage in the Parkinson's disease brain

Fiona Dick, Gonzalo S. Nido, Guido Werner Alves, Ole-Bjørn Tysnes, Gry Hilde Nilsen, Christian Dölle, Charalampos Tzoulis
Plos Genetics, **16/11** (2020)

PLOS GENETICS

RESEARCH ARTICLE

Differential transcript usage in the Parkinson's disease brain

Fiona Dick[1,2], Gonzalo S. Nido[1,2], Guido Werner Alves[3,4], Ole-Bjørn Tysnes[1,2], Gry Hilde Nilsen[1,2], Christian Dölle[1,2], Charalampos Tzoulis[1,2]*

1 Neuro-SysMed, Department of Neurology, Haukeland University Hospital, Bergen, Norway, **2** Department of Clinical Medicine, University of Bergen, Bergen, Norway, **3** The Norwegian Center for Movement Disorders and Department of Neurology, Stavanger University Hospital, Stavanger, Norway, **4** Department of Mathematics and Natural Sciences, University of Stavanger, Stavanger, Norway

* charalampos.tzoulis@uib.no

OPEN ACCESS

Citation: Dick F, Nido GS, Alves GW, Tysnes O-B, Nilsen GH, Dölle C, et al. (2020) Differential transcript usage in the Parkinson's disease brain. PLoS Genet 16(11): e1009182. https://doi.org/10.1371/journal.pgen.1009182

Editor: Bruce A. Hamilton, University of California San Diego, UNITED STATES

Received: May 28, 2020

Accepted: October 8, 2020

Published: November 2, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1009182

Data Availability Statement: The datasets supporting the conclusions of this article are included within the article and its supplementary files. The source code and raw data to reproduce the results of the analyses is available in the GitHub

Abstract

Studies of differential gene expression have identified several molecular signatures and pathways associated with Parkinson's disease (PD). The role of isoform switches and differential transcript usage (DTU) remains, however, unexplored. Here, we report the first genome-wide study of DTU in PD. We performed RNA sequencing following ribosomal RNA depletion in prefrontal cortex samples of 49 individuals from two independent case-control cohorts. DTU was assessed using two transcript-count based approaches, implemented in the DRIMSeq and DEXSeq tools. Multiple PD-associated DTU events were detected in each cohort, of which 23 DTU events in 19 genes replicated across both patient cohorts. For several of these, including *THEM5*, *SLC16A1* and *BCHE*, DTU was predicted to have substantial functional consequences, such as altered subcellular localization or switching to non-protein coding isoforms. Furthermore, genes with PD-associated DTU were enriched in functional pathways previously linked to PD, including reactive oxygen species generation and protein homeostasis. Importantly, the vast majority of genes exhibiting DTU were not differentially expressed at the gene-level and were therefore not identified by conventional differential gene expression analysis. Our findings provide the first insight into the DTU landscape of PD and identify novel disease-associated genes. Moreover, we show that DTU may have important functional consequences in the PD brain, since it is predicted to alter the functional composition of the proteome. Based on these results, we propose that DTU analysis is an essential complement to differential gene expression studies in order to provide a more accurate and complete picture of disease-associated transcriptomic alterations.

Author summary

Altered expression has been found at the level of genes and pathways in the brain of individuals with Parkinson's disease but remains unexplored at the level of individual transcripts. Thus, it is largely unknown whether transcript-specific events, for instance due to altered splicing or post-transcriptional modifications, occur in the Parkinson's disease

repository "DTUinPDbrain", https://github.com/fifdick/DTUinPDbrain under the GPL public license v3.0. Result tables, raw counts (in TPM) and sample information used to run the analyses are available at Figshare: https://figshare.com/articles/dataset/Differential_transcript_usage_in_the_Parkinson_s_disease_brain/12941945 Raw FASTA files include sensitive information and can not be shared publicly.

Funding: This work is supported by grants from The Research Council of Norway (288164, ES633272) (https://www.forskningsradet.no/en/) and Bergen Research Foundation (BFS2017REK05) (https://mohnfoundation.no/engelsk-rekruttering/?lang=en). Both of these were received by CT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

brain. Using RNA sequencing data from 49 brain samples, we performed a transcriptome-wide study of differential transcript usage in Parkinson's disease. We identified transcript-specific changes in multiple genes, and many of these were predicted to have important functional consequences on the encoded protein, such as altered subcellular localization or total protein levels. Interestingly, the vast majority of these transcript-specific changes were not detected by conventional differential gene expression analysis. Our findings suggest that analyses of differential transcript usage can provide additional insight into the transcriptomic landscape of complex brain disorders.

Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder, affecting more than 1% of the population above the age of 60 years [1]. Both genetic and environmental factors influence the risk of PD, but the molecular mechanisms underlying disease initiation and progression remain unknown. Studies of differential gene expression (DGE) employing microarrays or RNA sequencing (RNA-Seq) have identified molecular signatures associated with PD, including various aspects of mitochondrial function, protein degradation, neuroinflammation, vesicular transport and synaptic transmission [2].

An important limitation of DGE studies, however, is that they do not account for isoform diversity. Most genes encode more than one transcript isoform (henceforth called isoform), arising from alternative splicing, alternative usage of transcription start sites, or post-transcriptional regulation events such as alternative cleavage and polyadenylation [3]. Distinguishing between isoforms is essential, as these can encode proteins with different functions and/or subcellular localizations, or no protein product at all. Isoforms can also be associated with varying degrees of mRNA stability, for example by varying the length of the 3'-untranslated regions, which ultimately influences the rate of translation and hence the quantity of the encoded protein [4]. Moreover, differential splicing can impact cellular function without causing major changes on the levels of expressed protein. The diversity of tissue-specific isoform expression patterns is mainly attributed to differential usage of untranslated transcripts and/or non-principal isoforms, suggesting that even small changes in isoform usage can have a substantial effect on the composition and function of the proteome [5].

An efficient method to characterize differences in the isoform landscape is via differential transcript usage (DTU) analysis. DTU is a measure of the relative contribution of one transcript to the overall expression of the gene (i.e. the total transcriptional output). The analysis is based on individual transcript read counts normalized to the sum of all transcript read counts of the gene. This sets DTU apart from differential transcript expression (DTE), where the individual transcript counts are investigated independently from the context of the total transcriptional output. DTU requires at least one DTE event for the usage ratio between the transcripts of a gene to change. In contrast, DTE can occur without DTU, when the expression of an isoform is altered but its relative contribution to the total transcriptional output remains unchanged [6].

Individual transcript-level information—DTE or DTU—is lost in conventional DGE analysis, where the counts of individual transcripts are collapsed at the gene level. DTU events changing in opposite directions (e.g. when one transcript is up-regulated and another down-regulated) may cancel out at the gene level. Thus, transcript usage quantification has the potential to identify candidate genes and processes which would otherwise remain concealed in traditional DGE and DTE studies.

In the human brain, specific transcript usage profiles have been associated with neuronal development and aging [7] as well as with disease [8], including neurodegeneration [9, 10]. Current evidence suggests that differential splicing and DTU may be implicated in PD [11]. Disease-associated alternative splicing has been reported for genes linked to idiopathic and monogenic PD, including *SNCA* [12], *PRKN* [12, 13] and *PARK7* [14]. With the exception of these targeted, hypothesis-based studies, however, the role of DTU in PD remains largely unexplored and no genome-wide DTU studies have been carried out to date.

In the present study we report the first genome-wide analysis of DTU in PD. We show that DTU does occur in the PD brain and identify genes that show robust, altered isoform ratios across two separate cohorts of individuals with idiopathic PD and neurologically healthy controls: a discovery cohort from the Park West study [15] (n = 28) and a replication cohort from the Netherlands Brain Bank (n = 21).

Results

Multiple DTU events are detected in the PD prefrontal cortex

We first analyzed RNA-Seq data from the prefrontal cortex of our discovery cohort (n = 17/11 PD/controls; Table A in S1 File), using two alternative approaches (DRIMSeq [16] and DEXSeq [17]) to characterize DTU between PD and controls. Statistically significant DTU surviving multiple testing correction are referred to as DTU events and a gene exhibiting at least one DTU event is referred to as a DTU gene (detailed definitions are provided in the Methods).

In the discovery cohort, DTU analysis was based on n = 40, 520 transcripts and identified 814 DTU events in 584 DTU genes. The analysis with DEXSeq identified 254 DTU genes and 495 DTU genes were reported by DRIMSeq, with 165 detected by both methods (Fig 1A). The number of single DTU events per DTU gene ranged from one to three (Table 1). The most common Ensembl transcript biotype involved in DTU events was "protein coding" for both DEXSeq and DRIMSeq, followed by "processed transcript" (i.e., transcripts not containing an ORF) and "retained intron" (i.e., transcripts containing intronic sequences) (Fig 1B). We tested for overrepresentation of DTU events across transcript biotypes using Fisher's exact test and found that DTU events were overrepresented in 3 categories for DRIMSeq after multiple testing correction at alpha 0.05 (protein coding, retained intron, antisense). Although no categories were significantly overrepresented after Bonferroni correction using DEXSeq, the lowest p-values were for "antisense" and "protein coding", in agreement with DRIMSeq. Test statistics for each of the biotype categories are listed in S1 Table.

Visualization of the overall behavior of the effect size as a function of the mean transcript expression (MA-plot) and nominal transcript significance (Volcano-plot) are shown in S1A and S1B Fig. The p-value distribution varied depending on the number of transcripts a gene possessed. This variation behaved differently in DRIMSeq and DEXSeq—the p-value distribution became more uneven with increasing numbers of transcripts in DRIMSeq and decreasing number of transcripts in DEXSeq (S2C Fig). A list of identified DTU events is provided in Table B in S1 File.

Gene-set enrichment analysis (GSEA) of the DTU genes showed clusters of enriched pathways related to regulation of cell development, identical protein binding and perinuclear region of cytoplasm as the top most significant in each of the GO Ontology categories (Biological process, Molecular function, Cellular component) (Table 2).

To validate our methodology, we sought to confirm relative transcript abundances of genes with a DTU event by quantitative PCR (qPCR). To this end, we selected two genes fulfilling the following criteria: i) adequate individual transcript expression levels (i.e., the transcript

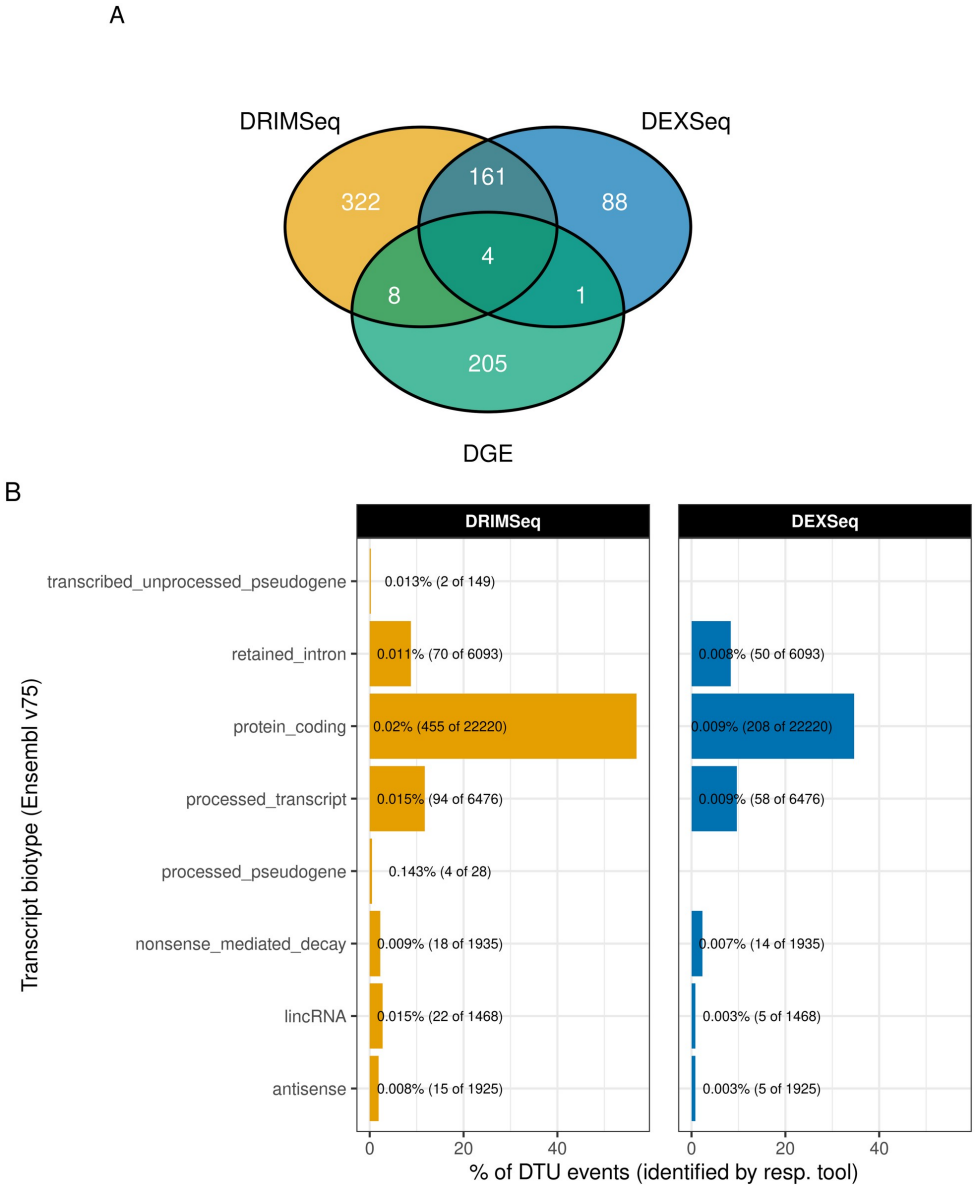


Fig 1. Overlap of DTU genes and transcripts between DEXSeq and DRIMSeq. A: Venn diagram showing the overlap between DTU genes resulting from analyses using DEXSeq and DRIMSeq, and genes that show DGE in the discovery cohort. B: Distribution of DTU events across defined transcript biotypes for each of the two tools (panels). Transcript biotypes are arranged on the y-axis, with the percentage of DTU events in each biotype category of all tool-specific DTU events represented on the x-axis. Text labels show the percentage of DTU events relative to the number of transcripts tested in each biotype category.

https://doi.org/10.1371/journal.pgen.1009182.g001

Table 1. Distribution of the number of DTU events per gene.

Tool	1 transcript	2 transcripts	3 transcripts
DEXSeq	173	76	5
DRIMSeq	312	181	2

https://doi.org/10.1371/journal.pgen.1009182.t001

was present in both cohorts after pre-filtering and detectable by qPCR) and ii) sufficiently distinct exonic composition of the individual transcripts to allow transcript-specific amplification (i.e., it was possible to design individual primer pairs that would detect one specific transcript variant alone). The genes *ZNF189* and *BCHE* satisfied all criteria and their transcript variants could be successfully amplified, serving as a proof-of-principle target (Fig 2A). The qPCR

Table 2. Enriched GO pathway clusters.

	Pathway	p-value
GO biological process		
	regulation of cell development	$6.62 \cdot 10^{-14}$
	regulation of nitric oxide biosynthetic process	$1.96 \cdot 10^{-10}$
	mitotic cell cycle	$6.43 \cdot 10^{-10}$
	regulation of transport	$1.71 \cdot 10^{-08}$
	nuclear DNA replication	$1.20 \cdot 10^{-06}$
	regulation of cellular component size	$1.46 \cdot 10^{-06}$
	phosphate-containing compound metabolic process	$5.51 \cdot 10^{-06}$
	single-organism catabolic process	$5.65 \cdot 10^{-06}$
	negative regulation of transcription, DNA-templated	$2.45 \cdot 10^{-05}$
	neurotrophin TRK receptor signaling pathway	$3.52 \cdot 10^{-05}$
GO molecular functions		
	identical protein binding	$7.26 \cdot 10^{-16}$
	nucleic acid binding transcription factor activity	$2.15 \cdot 10^{-10}$
	ubiquitin-protein transferase activity	$1.78 \cdot 10^{-08}$
	protein kinase binding	$7.60 \cdot 10^{-08}$
	zinc ion binding	$6.15 \cdot 10^{-06}$
	substrate-specific transporter activity	$7.63 \cdot 10^{-05}$
	transcription cofactor activity	$8.21 \cdot 10^{-05}$
	protein serine/threonine kinase activity	$1.26 \cdot 10^{-03}$
	DNA-directed DNA polymerase activity3	$1.82 \cdot 10^{-03}$
	Ras guanyl-nucleotide exchange factor activity	$2.41 \cdot 10^{-03}$
GO cellular component		
	perinuclear region of cytoplasm	$2.32 \cdot 10^{-04}$
	nuclear speck	$9.90 \cdot 10^{-04}$
	nuclear chromosome part	$4.04 \cdot 10^{-03}$
	plasma membrane part	$1.33 \cdot 10^{-02}$
	intercellular bridge	$1.56 \cdot 10^{-02}$
	cell projection	$1.74 \cdot 10^{-02}$
	nuclear envelope	$1.95 \cdot 10^{-02}$
	nucleolus	$3.10 \cdot 10^{-02}$
	membrane protein complex	$3.18 \cdot 10^{-02}$

Displayed are the titles of each pathway cluster. A cluster consists of multiple pathways that share a set of genes and have shown high overlap. Only significant pathways after correction have been considered for the clustering. The list of clusters is sorted by the aggregated p-values of each pathway in one cluster.

https://doi.org/10.1371/journal.pgen.1009182.t002

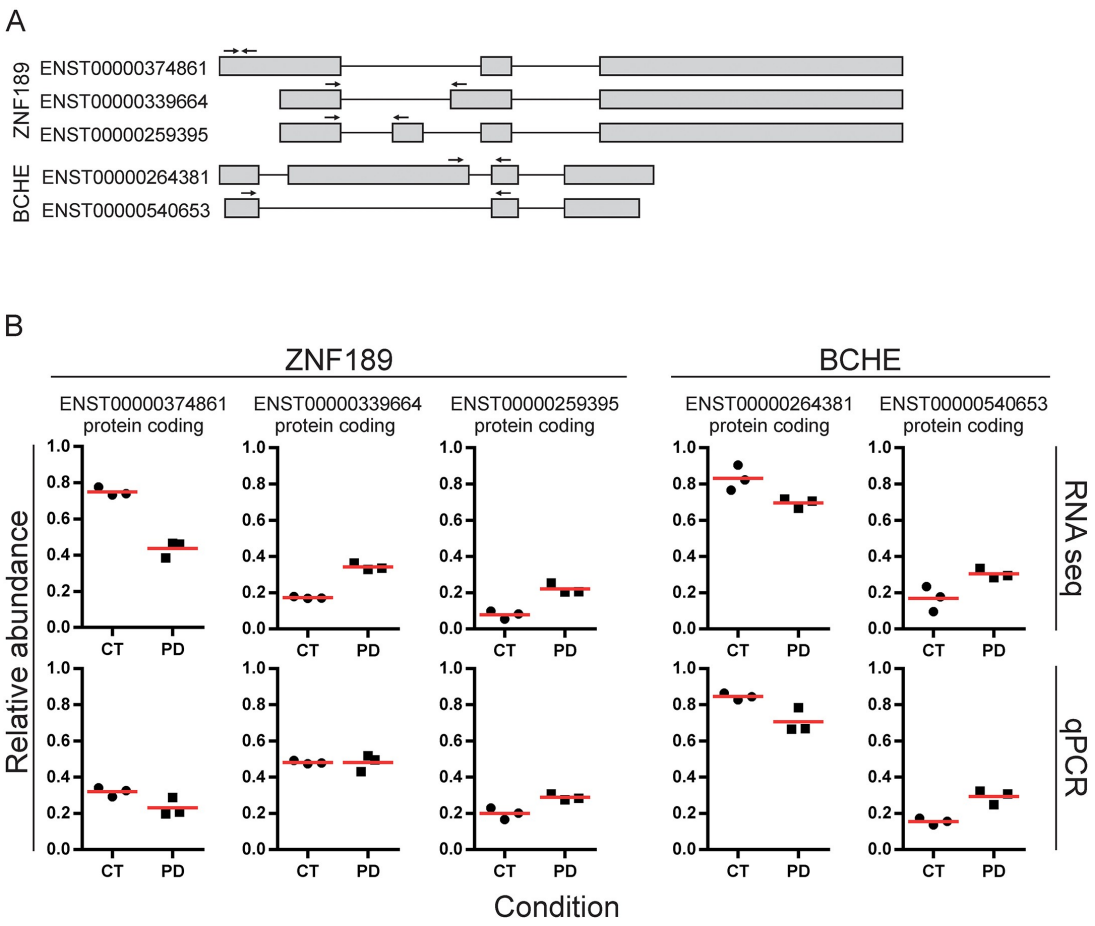


Fig 2. qPCR validation of *ZNF189* and *BCHE* relative transcript abundances in individuals with PD and controls. A: Schematic representation of ZNF189 and BCHE transcript variants analysed by qPCR. qPCR primer positions are indicated by arrows. B: Comparison of relative transcript abundances for the genes *ZNF189* and *BCHE*, obtained from RNASeq and qPCR. The upper row represents raw relative transcript abundances. Listed are only transcripts that remained after filtering. Data points are grouped by condition on the x-axis (PD vs CT). The three data points per group represent the three samples selected for qPCR. The lower row represents the results of qPCR analysis. Red lines show the mean of the respective group.

https://doi.org/10.1371/journal.pgen.1009182.g002

analysis replicated the results of the RNA-Seq-based DTU analyses for two of the three isoforms of *ZNF189* (ENST00000374861 and ENST00000259395), while the third isoform (ENST00000339664) appeared unchanged (Fig 2B). The qPCR analysis for *BCHE* confirmed the increased relative expression of isoform ENST00000540653 and the decreased relative expression of isform ENST00000264381 (Fig 2B).

Pre-filtering reduces transcriptome complexity

To reduce the false discovery rate (FDR), transcripts and genes underwent a pre-filtering based on a minimum expression level prior to the analysis (see Methods). This pre-filtering

affected the distribution of mean transcript expression and the mean number of transcripts per gene. In the discovery cohort, 77% (n = 137, 437) of all transcripts and 75% (n = 38, 100) of all genes were removed due to insufficient expression. Likewise, 82% (n = 143, 823) of all transcripts and 78% (n = 39, 342) of all genes were filtered out in the replication cohort. The distribution of mean transcript expression in the discovery cohort was shifted from a median of 15 read counts to 61, and from 12 to 63 in the replication cohort, after excluding low expressed transcripts and genes. The filtering procedure reduced the standard deviation of the mean transcript distribution in both cohorts from 30,753 to 432 in the discovery cohort, and from 39,096 to 484 in the replication cohort (Fig 3A). We also observed a reduction in the median number of transcripts per gene, from 9 to 3 in the discovery cohort and from 10 to 3 in the replication cohort (Fig 3B). We also observed an increase in the relative amount of protein coding transcripts as well as a decrease in the amount of pseudogene transcripts, snoRNAs, snRNAs, miRNAs and rRNAs (Fig 3C).

Alternative DTU methods agree in effect size and are minimally influenced by accounting for cell type composition

We investigated the agreement of effect size (i.e., the modeled coefficient for the disease state) in terms of magnitude and direction between the two tools in the discovery cohort. Overall, both methods agreed on the estimated effect size ($R = 0.97$, $p = 2.2 \cdot 10^{-16}$, $n = 40, 520$) and the concordance was even more pronounced in the subset of DTU events that were significant for either one of the cohorts ($R = 0.98$, $p = 2.2 \cdot 10^{-16}$, $n = 813$) (Fig 4). The general trend of statistical significance showed that transcripts which were identified as DTU events by at least one of the methods were likely to be defined at least as nominally significant by the alternative method: 97% of all DRIMSeq DTU events were nominally significant according to DEXSeq and 98% of all DEXSeq DTU events were reported as nominally significant by DRIMSeq. The concordance between the two methods in the replication cohort is shown in S2 Fig. We have recently shown that cell type heterogeneity can have a substantial impact on DGE analyses in bulk brain tissue [18]. To determine whether this also applied to our DTU analyses, we assessed the effect of accounting for cell type composition on our results. To this end, we obtained relative cellularity estimates (marker gene profiles, MGPs) for the cortical cell-types that were shown to be significantly associated with disease status (oligodendrocytes and microglia) in our previous study employing the same samples [18]. Accounting for cellular composition slightly increased the discovery signal, identifying a few more DTU genes with both DRIMSeq and DEXSeq. This effect was minor, however, as most DTU genes and events were identified irrespective of whether cell-type composition was accounted for or not (S3 and S4 Figs).

Most DTU events are not detected by conventional DGE analysis

Next, we sought to determine whether DTU events were detectable at the gene level by comparing the results of the DTU analysis to a conventional DGE analysis performed on the same dataset [18]. We found that less than 3% (n = 13) of the DTU genes (n = 584) were also significant at the gene level (BH corrected, $FDR < 0.05$) (Fig 1A), suggesting that compensatory changes across transcripts can balance out overall gene expression. Indeed, in genes with two DTU events, the effect size of these generally tended to move in opposite directions, canceling out the change in overall gene expression (Fig 5A). Similarly, in genes with only one DTU event, the effect size of DGE was smaller than the effect size of DTU, or even close to zero (Fig 5B), which likely originated from compensation distributed across multiple transcripts.

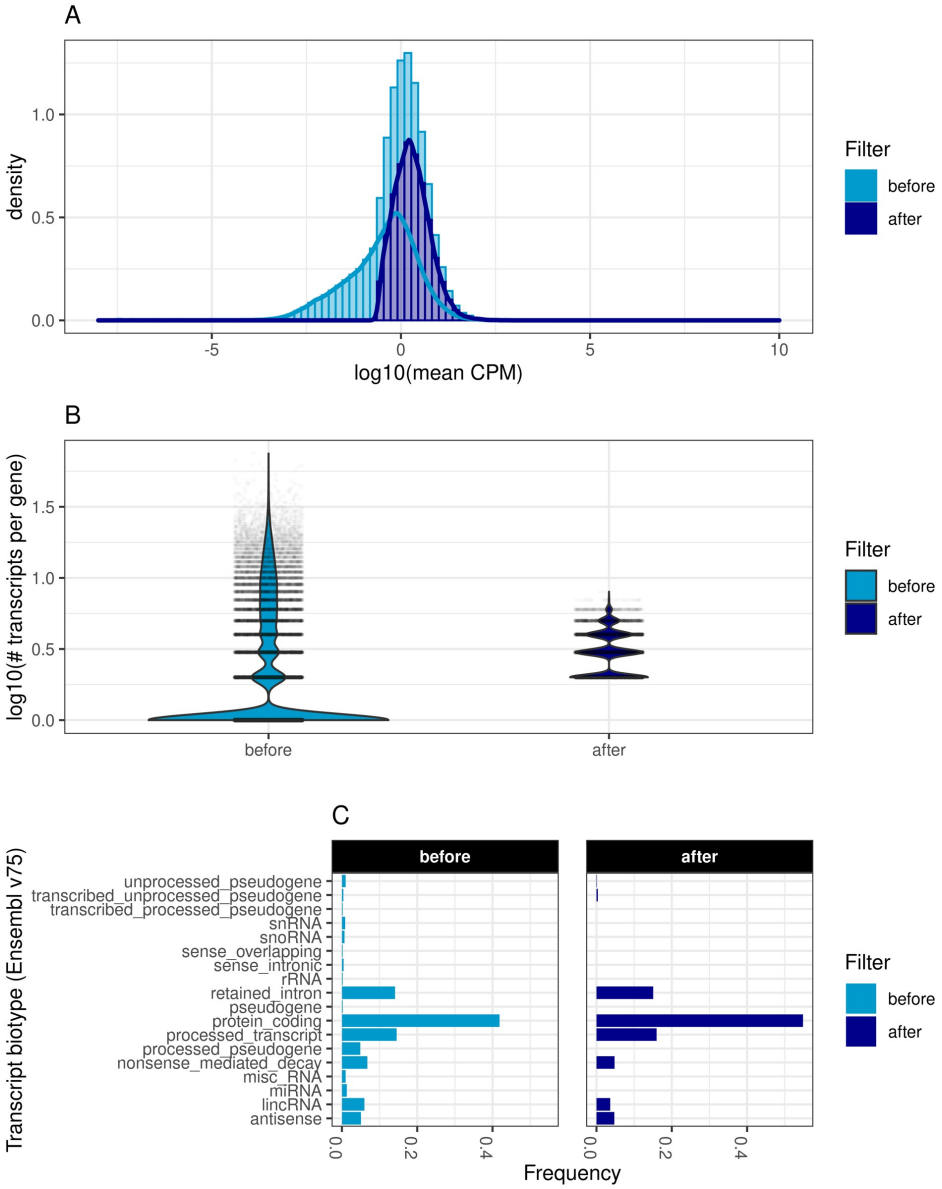


Fig 3. Transcript filter statistics. Comparison of data distributions before and after filtering of low expressed transcripts and genes. A: Log10 of the mean CPMs (counts per million) over all samples before and after filtering out low-expressed transcripts and genes. B: Violin plots showing the distribution of the number of transcripts per gene (in logarithmic scale). Violin width is scaled by the total number of observations while jittered points represent actual observations. C: Bar plot of transcript biotypes as defined by Ensemble v75 before and after filtering. Displayed are the relative frequencies of each category normalized by the number of transcripts before and after filtering. Categories with frequencies smaller than 0.001 were excluded for better visualization.

https://doi.org/10.1371/journal.pgen.1009182.g003

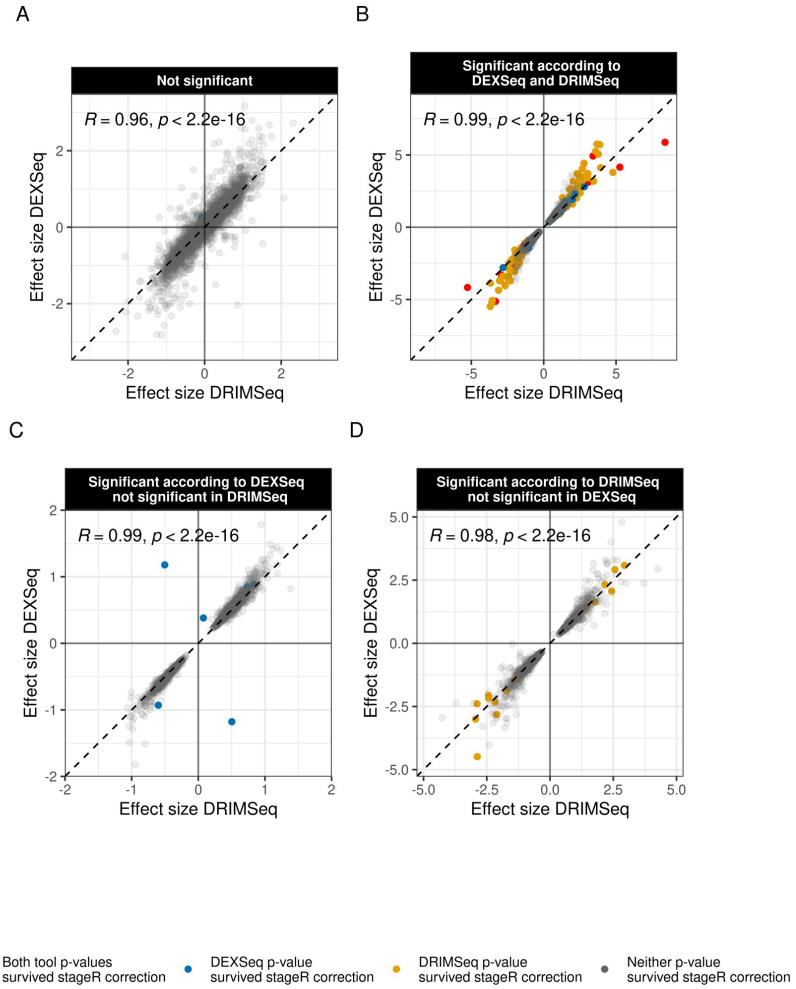


Fig 4. Concordance between DEXSeq and DRIMSeq. Estimated transcript usage effect sizes are shown for each transcript of the discovery cohort, with results from each tool on each of the axes (DRIMSeq x-axis, DEXSeq y-axis). Points situated on the diagonal represent transcripts with equal effect size in both tools. Points situated inside the first and third quadrant represent transcripts agreeing in direction across tools (i.e. first quadrant: up-regulated in PD, third quadrant: down-regulated in PD). A: Transcripts that did not reach statistical significance in the DTU analyses by either DRIMSeq or DEXSeq. B: Transcripts found to be significant by both tools. C: Transcripts found to be significant by DEXSeq only. D: Transcripts found to be significant by DRIMSeq only. Transcripts identified as DTU events (significant after p-value adjustment) are coloured according to the plot legend. Red: DTU event by both tools, blue: DTU event by DEXSeq only, yellow: DTU event by DRIMSeq only, grey: transcript either did not survive FWER correction in any of the tools or was not nominally significant.

https://doi.org/10.1371/journal.pgen.1009182.g004

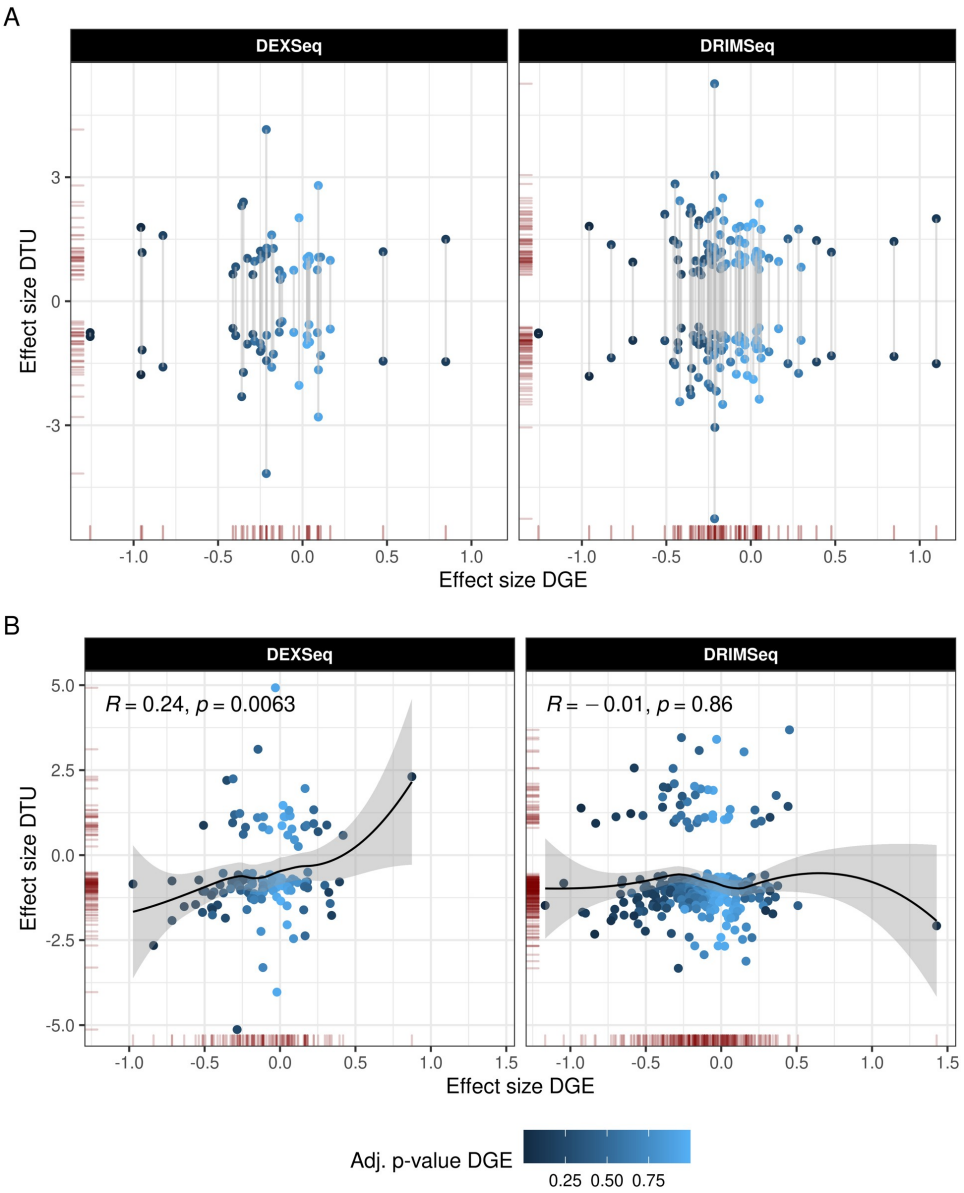


Fig 5. Concordance with DGE. The plot shows the relationship between the DTU effect size for each transcript (y-axis) and DGE effect size (x-axis). Data points correspond to transcripts. The x-coordinate of each point represents the effect size estimated for its parent gene in DGE analysis. The color scale indicates DGE significance after correction. A: DTU genes with 2 DTU events, with connected points representing each of the events from the gene. B: DTU genes with a single DTU event.

https://doi.org/10.1371/journal.pgen.1009182.g005

Table 3. DTU genes detected by DGE.

Tool	Gene	Transcript ID	Biotype	ES DTU	ES DGE
DRIMSeq	BCHE	ENST00000540653	protein coding	1.81	-0.96
DRIMSeq	BCHE	ENST00000264381	protein coding	-1.81	-0.96
DEXSeq	DAAM2	ENST00000491083	processed transcript	0.99	-0.80
DRIMSeq	EAF1-AS1	ENST00000610011	antisense	2.79	0.99
DRIMSeq	FRG1B	ENST00000439954	protein coding	-0.77	-1.26
DRIMSeq	FRG1B	ENST00000358464	protein coding	-0.79	-1.26
DRIMSeq	FRG1B	ENST00000479318	nonsense mediated decay	1.13	-1.26
DRIMSeq	FRMPD2	ENST00000491130	retained intron	-2.97	-1.17
DRIMSeq	FRMPD2	ENST00000486151	retained intron	2.97	-1.17
DRIMSeq	HIBCH	ENST00000414928	nonsense mediated decay	1.58	0.88
DRIMSeq	MIA	ENST00000597600	protein coding	-2.08	1.43
DRIMSeq	MIA	ENST00000593317	retained intron	2.08	1.43
DRIMSeq	MMP24-AS1	ENST00000566203	antisense	0.91	-0.58
DRIMSeq	PRODH	ENST00000334029	protein coding	-1.48	-1.17
DRIMSeq	SLCO1A2	ENST00000452078	protein coding	-0.83	-1.04
DRIMSeq	SLCO1A2	ENST00000463718	retained intron	0.83	-1.04
DRIMSeq	TSPAN15	ENST00000475069	retained intron	2.32	-0.84
DRIMSeq	TSPAN15	ENST00000373290	protein coding	-2.32	-0.84
DRIMSeq	UFSP2	ENST00000509180	protein coding	1.22	-0.60
DRIMSeq	VWF	ENST00000538635	processed transcript	-1.38	-0.93
DRIMSeq	VWF	ENST00000261405	protein coding	1.38	-0.93

Each transcript is described by its Ensemble identifier (version 75). The effect size (ES) is relative to the controls, i.e. positive ES represents an increase in PD relative to controls, negative ES a decrease. All entries in the table represent DTU events of which the parent gene was detected by DGE (BH adjusted, FDR = 0.05). DTU events that were identified by both DRIMSeq and DEXSeq are listed only with the estimated ES of DRIMSeq. The list is sorted by gene name in alphabetical order.

https://doi.org/10.1371/journal.pgen.1009182.t003

Only 13 DTU genes with at least one DTU event were also identified by DGE (Table 3). Six of these genes had a single DTU event and the remaining 7 had multiple DTU events. Of the 6 genes with a single DTU event, 3 showed the same direction of change in both DGE and DTU, whereas in the other 3, DGE and DTU indicated changes in opposite directions. For all 7 DTU genes with multiple DTU events, at least one DTU event was in the opposite direction of the DGE change. For example, while the protein coding transcript of the *VWF* gene was up-regulated, DGE analysis showed down-regulation at the gene-level, driven by a non-protein coding isoform. These results indicate that DTU analyses provide important additional insight into the transcriptomic landscape of PD.

Detected DTU events replicated in an independent patient cohort

We replicated our findings using RNA-Seq data from an independent cohort from the Netherlands Brain Bank (n = 10/11 PD/controls; Table A in S1 File). A total of 32,040 transcripts passed quality filtering in the replication cohort. The majority of these (n = 29, 807; 93%) overlapped with the pre-filtered transcripts of the discovery cohort and were further analyzed for replication. A total of 10,713 transcripts from the discovery cohort, however, did not pass pre-filtering in the replication cohort. Of these, 249 were identified as DTU events in the discovery cohort (S5A Fig). To assess the overall concordance between the two cohorts, we divided the common set of transcripts into 4 categories according to their nominal significance in differential usage in PD: i. non-significant in either cohort, ii. significant only in the discovery cohort, iii. significant in both cohorts, iv. significant only in the replication cohort. For each

category we assessed the concordance in DTU direction between the discovery and replication cohort (Fig 6A). In the group of non-significant transcripts, we observed a low correlation in the direction of DTU (Pearson's $R = 0.07$, $p = 2.2 \cdot 10^{-16}$, $n = 2, 5002$), with only 54% of transcripts agreeing between the cohorts. A higher correlation (Pearson's $R = 0.19$, $p = 2.2 \cdot 10^{-16}$, $n = 3776$) was observed for the group of transcripts which were nominally significant in the discovery cohort only, where 59% of transcripts showed the same direction of change in both cohorts. Transcripts which were significant only in the replication cohort showed no correlation (Pearson's $R = 0.058$, $p = 0.092$, $n = 843$) in the direction of DTU. The highest correlation (Pearson's $R = 0.25$, $p = 0.6 \cdot 10^{-3}$, $n = 186$) was observed in the group of transcripts that were nominally significant in both cohorts, with a 62% concordance in direction.

When we reduced the collection of transcripts to DTU events detected in the discovery cohort, we saw a high correlation (Pearson's $R = 0.28$, $n = 481$, $p = 2.5 \cdot 10^{-10}$), with 64% of these transcripts agreeing on the direction of change. This suggests that highly significant DTU events identified in our discovery cohort show a similar trend in our replication cohort (Fig 6B). Notably, 23% of the DTU genes identified in the discovery cohort were filtered out during pre-processing of the replication cohort and thus were excluded from this analysis.

A total of 23 DTU events in 19 genes detected in the discovery cohort were concordant in direction of change and nominally significant in the replication cohort (Table 4).

Among the 19 replicated DTU genes, 15 showed one DTU event and four comprised two DTU events per gene. Interestingly, in the four genes exhibiting two DTU events (*LINC00499*, *BCHE*, *THEM5*, *SLC16A1*), these moved in opposite directions. In *BCHE* and *THEM5*, DTU resulted in isoform switches (i.e. two DTU events in opposite directions) between different protein-coding transcripts. *THEM5*, encoding an acyl-CoA thioesterase involved in mitochondrial fatty acid metabolism, showed decreased usage of the full-length transcript (encoding a 247 amino acid protein) and increased usage of a shorter transcript (encoding a 119 amino acid protein) in PD. The down-regulated, full-length isoform was predicted to localize to the mitochondria (*likelihood* = 0.99), whereas the up-regulated, shorter isoform was more likely to localize to the extracellular space (*likelihood* = 0.36) than to the mitochondria (*likelihood* = 0.21). Hence, the decreased usage of the full-length isoform could result in a decrease of mitochondrial *THEM5* activity in PD. A similar pattern was observed for the *BCHE* gene, encoding a butyrylcholinesterase, with the full-length isoform (encoding a protein of 602 amino acids) down-regulated in PD, and an up-regulated shorter transcript encoding a putative protein of 64 amino acids. While both isoforms were predicted to be soluble and localize to the extracellular space, the shorter isoform lacks the substrate binding site located at positions 144 and 145 and it is therefore predicted to be non-functional, suggesting that *BCHE* function may be down-regulated in PD. The *SLC16A1* gene, encoding a lactate transporter in oligodendroglia, showed a switch from a protein-coding to a non-protein coding isoform in PD, revealing decreased expression of the protein coding transcript in PD.

In agreement with the down-regulation observed at the gene level, only 2 out of 19 replicated genes with DTU showed a significant altered overall gene expression: *BCHE* and *PRODH* (BH corrected, $FDR < 0.05$). In the case of *BCHE*, the down-regulation was observed for the full-length transcript as described above. *PRODH* exhibited a single DTU event consisting of a decreased relative expression of a protein-coding transcript variant in PD.

No evidence of DTU for genes linked to monogenic PD

Previous research had suggested that genes linked to monogenic PD, including *SNCA*, *PARK7* and *PRKN*, may exhibit altered transcript expression patterns in idiopathic PD [11, 12, 14]. Therefore, we sought to investigate whether these observations replicate in our data.

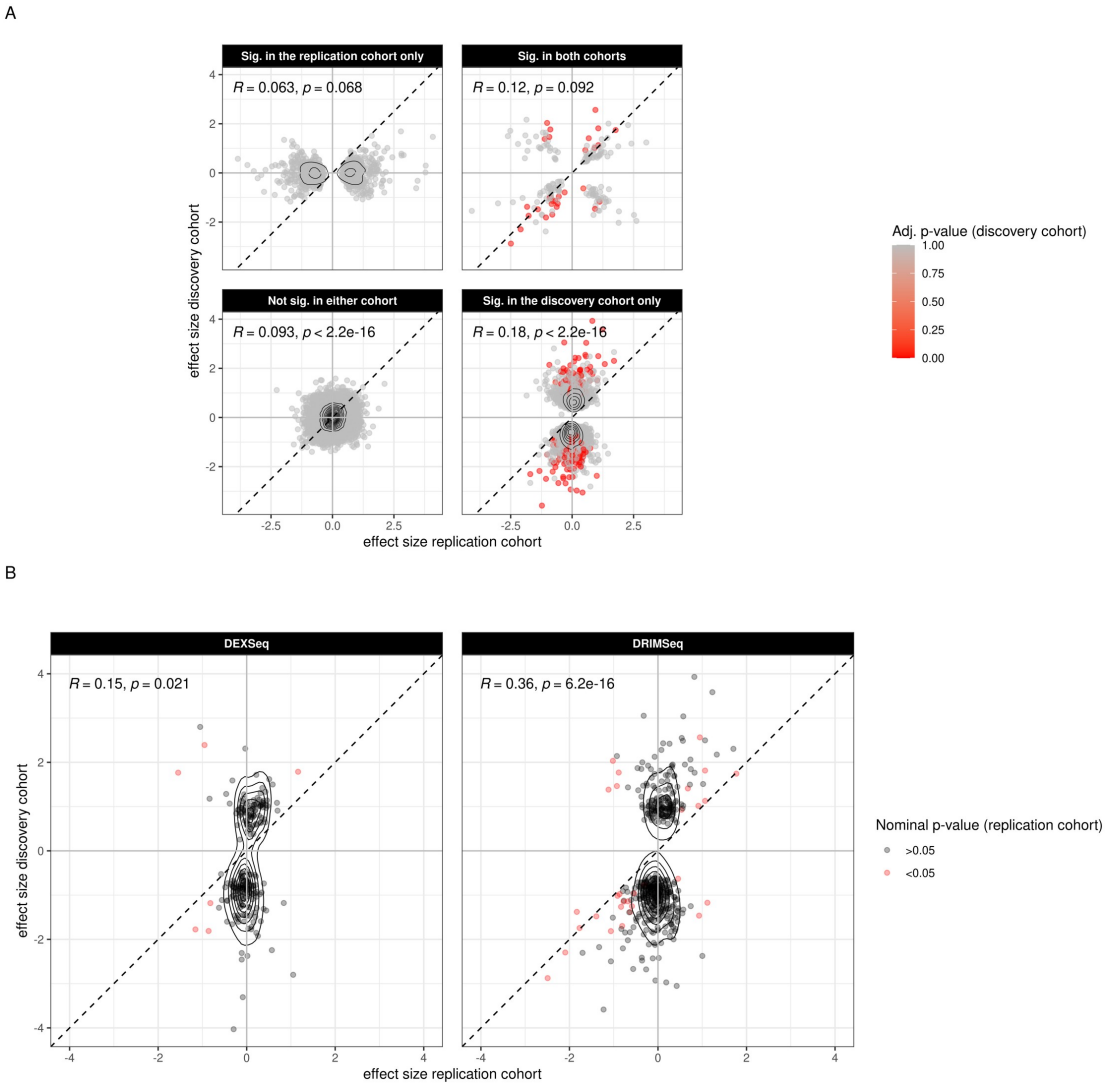


Fig 6. DTU replication in an independent cohort. Each data point corresponds to one transcript. The estimated effect size in the discovery cohort is represented on the y-axis and the estimated effect size for the replication cohort on the x-axis. A: The overlapping set of transcripts between the two cohorts is divided into 4 categories depending on their nominal significance in either cohort according to DRIMSeq. Transcripts not found to be significant in any cohort are shown in the lower left quadrant. Transcripts found to be significant in both cohorts in the upper right quadrant. Transcripts found to be significant only in the replication or in the discovery cohort are in the upper left and lower right quadrant respectively. The color scale (red-gray) shows adjusted p-value in the discovery cohort. B: Displayed are all DTU events (significant after correction) of the discovery cohort. Red color indicates nominal significance (before correction) in the replication cohort. The two columns present the results obtained from each respective tool.

https://doi.org/10.1371/journal.pgen.1009182.g006

Table 4. Replicated DTU genes.

Tool	Gene	Transcript ID	Biotype	ES discovery cohort	ES replication cohort
DEXSeq	BCHE	ENST00000264381	protein coding	-1.78	-1.16
DEXSeq	BCHE	ENST00000540653	protein coding	1.79	1.16
DEXSeq	XPA	ENST00000375128	protein coding	-1.18	-0.82
DEXSeq	VWA9	ENST00000573314	nonsense mediated decay	-1.81	-0.86
DRIMSeq	SLC16A1	ENST00000369626	protein coding	-1.02	-0.91
DRIMSeq	SLC16A1	ENST00000478835	processed transcript	1.02	0.91
DRIMSeq	THEM5	ENST00000453881	protein coding	1.74	1.77
DRIMSeq	THEM5	ENST00000368817	protein coding	-1.74	-1.77
DRIMSeq	BCHE	ENST00000540653	protein coding	1.81	1.06
DRIMSeq	BCHE	ENST00000264381	protein coding	-1.81	-1.06
DRIMSeq	HDAC3	ENST00000305264	protein coding	-1.38	-0.63
DRIMSeq	XPA	ENST00000375128	protein coding	-1.15	-0.78
DRIMSeq	ZNF208	ENST00000601993	protein coding	1.13	1.07
DRIMSeq	VWA9	ENST00000573314	nonsense mediated decay	-1.70	-0.81
DRIMSeq	SLC2A4RG	ENST00000473157	processed transcript	-2.30	-2.09
DRIMSeq	CD46	ENST00000367041	protein coding	-1.25	-0.59
DRIMSeq	ST3GAL5	ENST00000393808	protein coding	-0.96	-0.54
DRIMSeq	ACO1	ENST00000379923	protein coding	-1.38	-1.83
DRIMSeq	PRODH	ENST00000334029	protein coding	-1.48	-1.39
DRIMSeq	RPS9	ENST00000391752	protein coding	2.56	0.95
DRIMSeq	LRTOMT	ENST00000440313	protein coding	1.41	0.68
DRIMSeq	LINC00499	ENST00000510736	lincRNA	-1.14	-0.68
DRIMSeq	RNF38	ENST00000377885	protein coding	-0.98	-0.87
DRIMSeq	APIP	ENST00000527830	processed transcript	-2.87	-2.49
DRIMSeq	CNPY2	ENST00000548013	retained intron	-0.79	-0.31
DRIMSeq	LINC00499	ENST00000502757	lincRNA	0.93	0.54
DRIMSeq	ACAA1	ENST00000452171	protein coding	-1.26	-0.83

Each transcript is described by its Ensemble identifier (version 75). The effect size (ES) is relative to the controls, i.e. positive ES represents an increase in transcript usage in PD relative to controls, negative ES a decrease. The p-value as reported by stageR for each tool separately (DEXSeq and DRIMSeq) is representative for the level of significance after FWER control with $\alpha = 0.05$ and is lower than 0.03 for all listed DTU events. The table is sorted by the p-value in increasing order and grouped by the tool that identified the transcript

https://doi.org/10.1371/journal.pgen.1009182.t004

Increased expression of four *SNCA* transcript variants, encoding the protein isoforms SNCA-140, SNCA-126, SNCA-112 and SNCA-98, were reported in the prefrontal cortex of individuals with PD [12]. None of these transcripts showed evidence of DTU in our analysis. The transcript (ENST00000506244) encoding the full-length protein (SNCA-140), showed a trend for reduced relative expression in PD, but this did not reach statistical significance ($p = 0.055$, effect size = −0.48, DRIMSeq). In the same study, two out of seven protein-coding splice variants of PRKN (TV3 and TV12) were suggested to be overexpressed in the PD brain. In our data, only two PRKN transcript variants (TV1 and TV2) showed sufficient expression to be analyzed, and neither of them showed statistical evidence of DTU (nominal $p > 0.79$, absolute effect size < 0.09, DRIMSeq) in agreement with the results reported in [12].

Finally, one study reported that the altered relative transcript abundance of *PARK7* in blood may be used as a biomarker for PD [14]. None of the transcript variants of *PARK7* were sufficiently expressed in our dataset to investigate the transcript usage pattern of this gene in the PD brain.

Discussion

We report the first transcriptome-wide DTU study in PD. Our analyses reveal that multiple DTU events occur in the PD brain and many of these are predicted to have a functional impact. Interestingly, the vast majority of genes exhibiting DTU are not detected by conventional DGE analysis on the same dataset. This is either because DTU occurs in low-expressed isoforms, or due to antagonistic, inverse changes in other transcripts of the same gene, canceling out the net change at the gene expression level.

Our findings suggest that DTU events in PD may have important downstream consequences for protein function, irrespective of whether there is a measurable difference in the total gene expression levels. Changes in the relative expression of different transcripts of a gene affect the ratio of the resulting protein isoforms and could, therefore, influence biological processes through variation in function and/or subcellular localization. Moreover, switches may occur between protein coding and non-coding transcript isoforms, thereby affecting the overall protein level. Changes in the usage ratios of low expressed and/or non-protein coding isoforms may also have important biological effects, as it has been shown that these are highly cell- and tissue-specific, and have a substantial impact on the composition and function of the proteome [5].

In our dataset, individuals with PD showed a significant decrease in the relative usage of a *THEM5* transcript variant that encodes the full-length THEM5 protein isoform, predicted to localize to mitochondria. This isoform is involved in mitochondrial fatty acid metabolism by exhibiting esterase activity with a preference for long and unsaturated fatty acid-CoA esters [19]. Decreased *THEM5* function has been shown to influence the remodeling process of mitochondrial inner membrane cardiolipin [19, 20], resulting in abnormal mitochondrial morphology and impaired mitochondrial respiration [19], both of which occur in PD [18, 21]. A concomitant increase in the relative expression of a shorter THEM5 isoform resulted in relatively unchanged levels of total gene expression. However, as this isoform encodes a protein lacking the first 37 N-terminal amino acids, it is unlikely to localize to mitochondria, and may therefore not replace the full-length protein functionally [19].

A protein-coding transcript of the *SLC16A1* gene was significantly down-regulated in the PD brain and accompanied by an increase of similar magnitude in a non-protein coding transcript. *SLC16A1* encodes a monocarboxylate transporter (*MCT1*) responsible for lactate and pyruvate trafficking across cell membranes. *MCT1* is the most abundant lactate transporter in the central nervous system, where it is highly expressed in oligodendroglia. It has been shown that *MCT1* plays a key role in the energy homeostasis of neurons, by regulating lactate transport between oligodendroglia and axons. *MCT1* disruption causes axonal dysfunction and neurodegeneration in cell and animal models and MCT1 levels have been found to be decreased in patients and mouse models of ALS [22, 23].

Another gene of interest was *BCHE*, which showed a decreased usage of the protein-coding full-length transcript, suggesting that the level of the functional full length protein isoform may be decreased in PD. Interestingly, genetic variation in this gene has been associated with Alzheimer's disease [24], susceptibility to pesticide toxicity [25] and, more recently, with PD [26].

In the few genes that were detected by both DTU and DGE analysis, DTU provided additional functional insight. Since changes in the relative isoform expression can occur in opposite directions to the overall gene-level expression, transcript-level resolution is essential in order to predict the functional consequences of altered expression.

Our analyses did not confirm a previous report of altered transcript expression in the *SNCA* gene in the PD frontal cortex [12]. These findings were based on a small PD cohort

(n = 5) with no reported neuropathological confirmation of the diagnosis. The fact that the reported transcripts were confidently detected in our data but showed no evidence (or trend) of altered relative expression in either of our cohorts, suggests that this effect, if real, is not a general or common phenomenon in PD. Alternatively, the lack of replication may reflect different genetic backgrounds and environmental exposures in different populations (Spanish, Norwegian and Dutch). The *PRKN* transcripts TV3 and TV12, which were reported to show altered expression in PD in the same sample as *SNCA* [12, 13] did not show sufficient expression in our material to be confidently assessed for replication.

While most identified DTU genes in our results do not have a known role in PD, pathway analyses showed significant enrichment in clusters associated with the pathophysiology of PD, including reactive oxygen species (ROS) generation and protein degradation. These results confirm that our findings are related to the biology of PD and highlight DTU analyses as a complementary strategy to nominating novel disease candidate genes and processes.

A potential limitation in our study is posed by differences in cell-type composition between brain tissue of patients and controls. We have recently shown that this can be an important confounding factor in differential expression analysis of bulk brain tissue [18]. To mitigate this problem, we accounted for differences in cellularity across samples by including cell type estimates for specific cell types found to be significantly associated with disease status, as covariates in our model. Notably, correcting for cell-type composition had only a minor effect in our results, supporting the notion that most identified DTU events are not driven by differences in cellularity between PD and controls.

While our top DTU findings replicate across the two independent cohorts, suggesting these changes are robustly associated with PD, we nevertheless observe an overall low concordance between the cohorts. This most likely reflects a combination of biological and technical factors, including limited power due to the relatively small sizes of the cohorts, heterogeneous disease biology and cell-composition, population-specific and/or brain bank-specific effects, differences in the age and RIN ranges. Differences between the cohorts were also evident in the filtering results, whereby a larger number of transcripts in the replication cohort were filtered out in comparison to the discovery cohort, as summarized in S5A Fig. We hypothesized that this may be related to the overall higher RINs of the samples from the replication cohort. Transcripts which were detected in the discovery cohort but not in the replication cohort showed a negative correlation with RIN (S5B Fig), suggesting that lower RNA quality (reflected by lower RIN values) is associated with higher transcript counts due to an increase in non-specific alignments in degraded samples.

Further replication in larger samples will be required in order to confirm and further dissect the DTU landscape of the PD brain. Methodological limitations should also be considered. While DRIMSeq was designed specifically for DTU analysis and assesses the relationship of each transcript abundance relative to the total transcriptional output, it may have difficulties to correctly estimate the dispersion for genes with a large number of isoforms [16]. This can potentially lead to inaccurate transcript proportion estimations and increase the susceptibility to false positive results, as suggested by the p-value distributions. Conversely, DEXSeq cannot capture the transcript-gene relationship directly, which might explain its general lower sensitivity compared to DRIMSeq.

Conclusion

In conclusion, our findings provide the first insight into the DTU landscape of PD. We show that DTU is a prominent feature in the PD brain and may have important functional consequences by altering the structural and functional composition of the proteome. We therefore propose that

DTU analyses should be an essential component of transcriptomic studies, along with DGE analyses, because they provide additional insight into the transcriptomic landscape and allow a more accurate prediction of the functional consequences of detected changes in gene expression.

Methods

Cohorts

Fresh-frozen prefrontal cortex tissue (Brodmann area 9) was available from two independent cohorts. The discovery cohort comprised individuals with idiopathic PD (n = 17) from the Park West study, a prospective population-based cohort, which has been described in detail [15], and demographically matched controls (n = 11). Samples were collected and stored in our Brain Bank for Aging and Neurodegeneration. The replication cohort comprised individuals with idiopathic PD (n = 10) and demographically matched controls (n = 11) from the Netherlands Brain Bank. The details of the cohorts are summarized in Table A in S1 File.

Ethics statement

Ethical permission for these studies was obtained from our regional ethics committee "Regional Committee for Medical and Health Research Ethics": REK 2017/2082, 2010/1700, 131.04 (REC, https://rekportalen.no/). Written formal informed consent was obtained from all participants or their next of kin.

RNA sequencing

Total RNA was extracted from prefrontal cortex tissue homogenate for all samples using RNeasy plus mini kit (Qiagen) with on-column DNase treatment according to manufacturer's protocol. Final elution was made in 65 μl of dH2O. The concentration and integrity of the total RNA was estimated by Ribogreen assay (Thermo Fisher Scientific), and Fragment Analyzer (Advanced Analytical), respectively. Five hundred ng of total RNA was required for proceeding to downstream RNA-seq applications. First, ribosomal RNA (rRNA) was removed using Ribo-Zero™ Gold (Epidemiology) kit (Illumina, San Diego, CA) using manufacturer's recommended protocol. Immediately after the rRNA removal the RNA was fragmented and primed for the first strand synthesis using the NEBNext First Strand synthesis module (New England BioLabs Inc., Ipswich, MA). Directional second strand synthesis was performed using NEBNext Ultra Directional second strand synthesis kit. Following this the samples were taken into standard library preparation protocol using NEBNext DNA Library Prep Master Mix Set for Illumina with slight modifications. Briefly, end-repair was done followed by poly(A) addition and custom adapter ligation. Post-ligated materials were individually barcoded with unique in-house Genomic Services Lab (GSL) primers and amplified through 12 cycles of PCR. Library quantity was assessed by Picogreen Assay (Thermo Fisher Scientific), and the library quality was estimated by utilizing a DNA High Sense chip on a Caliper Gx (Perkin Elmer). Accurate quantification of the final libraries for sequencing applications was determined using the qPCR-based KAPA Biosystems Library Quantification kit (Kapa Biosystems, Inc.). Each library was diluted to a final concentration of 12.5 nM and pooled equimolar prior to clustering. 125 bp Paired-End (PE) sequencing was performed on an Illumina HiSeq2500 sequencer (Illumina, Inc.) at a target depth of 60 million reads per sample.

FASTQ files were trimmed using Trimmomatic [27] to remove potential Illumina adapters and low quality bases with the following parameters:

```
ILLUMINACLIP:truseq.fa:2:30:10
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15.
```

FASTQ files were assessed using fastQC [28] prior and following trimming.

Transcript quantification

We used Salmon [29] with the fragment-level GC bias correction option (`--gcBias`) and the appropriate option for the library type (`-l ISR`) to quantify transcript expression in pseudo-alignment mode, using the GRCh37 genome as a reference. X and Y chromosomes were excluded from the GRCh37 reference genome, restricting quantification to transcripts located on autosomes.

Transcripts per million (TPM) values obtained with Salmon were scaled using the R package *tximport* [30] with the scaling method `scaledTPM`, the favored scaling method for DTU [31].

DTU analyses and quality control

DTU analyses estimate transcript usage and detect changes in the relative contribution of a transcript to the overall expression of the gene. Transcript usage corresponds to the transcript-level expression counts of a transcript *i* normalized by the sum of counts of all transcripts of a gene *j*:

$$TU_{i,j} = \frac{t_i}{\sum_{k=1}^{n_j} t_k} \quad , \tag{1}$$

where n_j equals the number of transcripts of gene *j* and t_i is the expression count of transcript *i*. Hence, *differential* transcript usage describes a change in proportions between the groups (PD and controls).

For our analysis, we employed an alignment-free abundance estimation method [29], which enabled read quantification at the transcript level directly, as opposed to traditional read alignment methods that require bin or exon read counting and subsequent summarization to transcript level.

We performed DTU analysis between PD and controls using two alternative approaches implemented in the tools DRIMSeq [16] and DEXSeq [17]. While DEXSeq was designed for detecting differential exon usage, it is also suitable for assessing DTU by using estimated transcript abundances directly [6, 31, 32]. DRIMSeq was developed specifically for DTU analyses and is based on estimated transcript counts [16]. These methods assess alternative splicing by directly identifying transcripts that are differentially used, rather than detecting specific splice events. Both methods have shown comparable performance in benchmarks with simulated data [16, 31, 32]. A further advantage was that these tools allow for the inclusion of known covariates into the model design. DRIMSeq assumes a Dirichlet multinomial model for each gene and estimates a gene-wise precision parameter, whereas DEXSeq assumes a negative binomial distribution for counts of each transcript and estimates a transcript-wise dispersion parameter [31]. It is worth noting that DRIMSeq bases its analyses directly on the calculated transcript proportions, thereby modeling the correlation among transcripts in their parent-gene directly, whereas those correlations may not be accurately captured by DEXSeq, as it models each transcript separately and accounts for gene-transcript interaction with a covariate in its model design [31].

Due to the complexity of the human transcriptome in terms of diversity and number of transcripts per gene, DTU methodologies tend to exhibit a worse performance considering the false discovery rate (FDR) when compared to simpler organisms [6]. However, FDR can be reduced considerably if the collection of transcripts undergoes filtering prior to analysis [6]. Transcript filtering, in addition, alleviates the DRIMSeq-specific difficulty of capturing the full bandwidth

of transcript dispersion through the common gene-level dispersion estimate [16], which results otherwise in a decrease in performance for genes with increasing number of transcripts. We thus excluded lowly expressed transcripts with a soft filter, allowing for a certain percentage of all samples to have a transcript expression below the given threshold. This filtering methodology was chosen over hard filtering in order to avoid overlooking cases of DTU driven by lack of expression in one of the groups being compared, which would have been the case with a hard threshold filtering. Using the filtering method available in the DRIMSeq package, we excluded transcripts for which more than `n = min(#Controls, #PD)` samples did not reach 10 read counts or for which their relative contribution to the overall gene expression was smaller than one percent. In addition, we filtered out genes with less than 10 counts in any one sample. To investigate changes in transcript usage between PD and controls, the resulting filtered set of transcript-level counts were used as an input for both DEXSeq and DRIMSeq as recently suggested by [31]. Analyses were carried out independently on both cohorts.

Model design

Sources of variation in our data were identified using principal component analysis (PCA) at the gene-level. RNA integrity number (RIN) correlated highly with the first principal component, indicating that RNA quality represents a major source of variation in the expression data.

Relative cellular composition in our samples was obtained from our previous study [18] using marker gene profiles (MGPs) [33, 34]. In summary, an MGP was calculated for each of the main cortical cell types (neurons, oligodendrocytes, astrocytes, endothelial, and microglia) by performing a PCA on the log-transformed expression (in counts per million) of cell type-specific marker genes from the NeuroExpresso database [33] and extracting the first principal component. MGPs for oligodendrocyte and microglia showed a significant association with the disease status (controls vs PD) and were accounted for in the DTU models together with RIN, gender, and age.

To explore the effect of accounting for disease-associated MGPs in the DTU results, we compared the two alternative designs, with and without oligodendrocyte and microglia MGPs. Accounting for cellular composition slightly increased the discovery signal, identifying a few more DTU genes with both DRIMSeq and DEXSeq. This effect was minor, however, as most DTU genes and events were identified irrespective of whether cell-type composition was accounted for or not (S3 and S4 Figs).

Statistical testing

The results of the DTU analyses were further processed with StageR [35]. Gene-level aggregated p-values (q-values) as well as transcript-level p-values were passed to stageR for a two-stage screening of significance. For DEXSeq, nominal p-values of all transcripts of a gene were aggregated to a q-value and corrected using the function *perGeneQvalue*. For DRIMSeq, nominal p-values were already reported at the gene-level and further corrected within stageR using the Benjamini-Hochberg (BH) FDR procedure. To control the FWER, transcript-level significance was corrected within-gene, if the gene passed the first screening stage of stageR, with respect to the FDR controlled gene-level significance (q-value). Transcripts of genes which did not pass the first screening stage, were not further assessed for significance at the transcript-level. Nominal transcript-level p-values of both tools were adjusted within StageR using an adapted Holm-Shaffer family-wise error rate (FWER) correction method specifically designed for DTU analysis [35].

We define a transcript as a *DTU event*, if the FWER-controlled $p < \alpha$ with $\alpha = 0.05$. Similarly, we define as *DTU gene* any gene that exhibits at least one DTU event.

Similarly, we define $\alpha = 0.05$ for nominal significance.

DTU pathway enrichment analysis

To assess the enrichment of DTU genes in predefined functional gene sets (pathways), we employed the *enrichment* function of the stringDB R package [36]. DTU genes identified in our discovery cohort were used as hits and all genes surviving the filtering step during pre-processing were used as background. Enrichment was tested for pathways defined by the Genome Ontology (GO) [37, 38]. Each of the three GO categories (Biological Process, Molecular Function, Cellular Compartment) was tested separately. To reduce redundancy of the top most enriched pathways ($FDR < 0.05$), we performed a clustering in each of the three GO categories. Pathways were clustered by iteratively joining nearest neighbors based on pathway similarity, which we defined with the Cohen's kappa coefficient (κ). The similarity of newly formed clusters and unvisited neighbours was iteratively recalculated, until no two clusters' κ was higher than a chosen threshold of 0.4. Each cluster was given a representative title, chosen from the names of all the pathways in a cluster. The choice of the cluster title depended on the pathway size, pathway significance or chosen randomly if none of the previous criteria were sufficient. Finally, each pathway cluster was assigned a p-value by aggregating p-values of all cluster members with the Fisher method.

For specific cases of isoform switches between protein coding transcripts, we used the tool DeepLoc [39] to predict subcellular localization by retrieving the encoded amino acid sequence from the Ensembl release 75.

RNA extraction, cDNA synthesis and quantitative PCR analysis

RNA extraction was carried out using the RNeasy Lipid Tissue Mini Kit (QIAGEN 74804), starting with ca. 20 mg brain tissue from three individuals with PD and three controls. 500 ng total RNA were subjected to cDNA synthesis using the SuperScript IV VILO Master Mix with ezDNase Enzyme (Thermofisher Scientific 11766500). Experiments were carried out in triplicates starting with a new cDNA synthesis from aliquoted total RNA. For the SYBR Green quantitative PCR analysis, the PowerUp SYBR Green Master Mix (Thermofisher Scientific, A25776) was used with a thermal cycling of one cycle at 95˚C for 20s and 40 cycles at 95˚C for 3s and 60˚C for 30s on a StepOnePlus instrument (Thermofisher Scientific), and with the primers listed in Table 5.

Table 5. qPCR primer sequences.

Transcript ID	Primer name	Primer sequence
ENST00000374861	ZNF189_374861 fw	5'-TGGGGTTCGGGGTTGGGG-3'
ENST00000374861	ZNF189_374861 rv	5'-CGGTCACGACCCCAACAGC-3'
ENST00000339664	ZNF189_339664 fw	5'-GATGGCTTCCCCGAGCCC-3'
ENST00000339664	ZNF189_339664 rv	5'-ACACAGCCACATCCTCAAATG-3'
ENST00000259395	ZNF189_259395 fw	5'-GAGATGGCTTCCCCGAGCC-3'
ENST00000259395	ZNF189_259395 rv	5'-CTTATTTTCTCAGGCCGATTTATC-3'
ENST00000540653	BCHE_540653 fw	5'- GCAAACTTTGCCATCTTTGTTG-3'
ENST00000540653	BCHE_540653 rv	5'- CTTGTGCTATTGTTCTGAGTC-3'
ENST00000264381	BCHE_264381 fw	5'- AGATCCATAGTGAAACGGTGG-3'
ENST00000264381	BCHE_264381 rv	5'- CTTGTGCTATTGTTCTGAGTC-3'
	GAPDH	Assay ID Hs00266705_g1 (Thermofisher)

https://doi.org/10.1371/journal.pgen.1009182.t005

Supporting information

S1 Fig. Diagnostic plots. Data points in all plots represent one transcript, with coloring showing significant transcripts ($\alpha = 0.05$) in red. P-values (uncorrected) are displayed as ($-log10$(p-value)). A: Volcano plot displaying the effect size (as estimated by the respective tool) in the x-axis and the p-value on the y-axis. Triangles mark extreme p-value outliers that were adjusted to fit into the plot. B: MA plot visualizing a transcript's significance as a function of its mean expression over all samples. C: Density ridges display the distribution of gene-level significance ($-log10$(p-value)) per gene type, where genes are grouped according to the number of transcripts they have after filtering. The color gradient was applied to visualize the p-value scale. The vertical dashed line corresponds to a p-value of 0.05.
(TIFF)

S2 Fig. Concordance between DEXSeq and DRIMSeq in the replication cohort. Estimated transcript usage effect sizes are shown for each transcript of the replication cohort, with results from each tool on each of the axes (DRIMSeq x-axis, DEXSeq y-axis). Points situated on the diagonal represent transcripts with equal effect size estimations of both tools; points situated inside the first and third quadrant of the coordinate system represent transcripts agreeing in direction according to both tools (i.e. up-regulated in PD: first quadrant, down-regulated in PD: third quadrant). A: Transcripts that did not reach statistical significance in the DTU analyses by either DRIMSeq or DEXSeq. B Transcripts found to be significant by both tools. C: Transcripts found to be significant by DEXSeq only. D: Transcripts found to be significant by DRIMSeq only. Transcripts identified as DTU events (significant after p-value adjustment) are coloured according to the plot legend. Red: transcript identified as a DTU event by both tools, yellow: transcript identified as a DTU event by DRIMSeq only, grey: transcript either didn't survive FWER correction by neither tool or wasn't nominally significant beforehand. (Transcripts can appear significant after FWER control even if they weren't nominally significant, due to StageR assigning significance by relying on the assumption that if DTU is occurring in the gene (that is: the gene has passed the screening stage) and one of its transcripts is significant, the other must subsequently also take part in the DTU to compensate).
(TIFF)

S3 Fig. Overlap DTU genes and events, with and without cell correction. DTU genes (A, B) and events (C, D) resulting from the analysis which included cell type estimations (purple) are overlapped with the results of the analysis where differences in cell types were not taken into account (turquoise). Only DTU events which were identified in the discovery cohort and replicated in the independent replication cohort were considered for this plot. A: DTU genes identified by DRIMSeq. B: DTU genes identified by DEXSeq. C: DTU events identified by DRIMSeq. D: DTU events identified by DEXSeq.
(TIFF)

S4 Fig. Characteristics of the replicated DTU genes and events depicted as heatmaps. Replicated DTU events (significant after OFWER correction in the discovery cohort, agreeing on the direction of change across cohorts and nominally significant at alpha = 0.05 in the replication cohort) are arranged in the y-axis. A: transcript's adjusted p-value (white cells indicate adjusted p-value >= 0.05). B: Transcript's log fold change (white cells correspond to transcripts not identified as DTU events). C: Transcript's nominal (uncorrected) p-value. In all heatmaps, characteristics are grouped by model design (i.e. with ("Incl. MGPs") or without ("w/o MGPs") accounting for MGPs) and by tool (DRIMseq or DEXSeq).
(TIFF)

S5 Fig. Effect of pre-filtering on the number of transcripts per cohort. A: Venn diagram for the sets of transcripts which survived pre-filtering in each cohort. Number of transcripts that survived filtering in the replication cohort (green), in the discovery cohort (red), and number of transcripts identified as DTU events in the discovery cohort (blue). B: Distribution of the correlation coefficients between transcript abundance (TPM) and sample RIN for non-concordant transcripts (i.e. transcripts removed during the pre-filtering in the replication cohort, but not in the discovery cohort) and concordant transcripts (i.e. transcripts that survived pre-filtering in both cohorts).
(TIFF)

S1 File. A: Cohort demographic and experimental information. B: DTU events. Table of identified DTU events, grouped by cohort (replication, discovery) and tool (DRIMSeq, DEXSeq). Gene-level and transcript-level p-values as reported by stageR (after FWER correction). Effect size corresponds to the coefficient of the condition variable (Control, PD) in the analysis model.
(XLSX)

S1 Table. Overrepresentation analysis of DTU events in transcript biotypes. P-values and odds ratios were determined by Fisher's exact test. The contingency table was built up separating transcripts by whether or not they were identified as DTU events and whether they were defined as the biotype of interest (as defined by Ensembl version 75). The rows are grouped by the tool which identified the DTU event and sorted by increasing p-value of the Fisher's exact test.
(PDF)

Acknowledgments

We are grateful to patients and their families for participating in our research. We would also like to thank our colleagues at the Neuromics group for the fruitful discussions.

Author Contributions

Conceptualization: Fiona Dick, Gonzalo S. Nido, Christian Dölle, Charalampos Tzoulis.

Data curation: Fiona Dick, Gonzalo S. Nido.

Formal analysis: Fiona Dick, Gonzalo S. Nido, Christian Dölle, Charalampos Tzoulis.

Methodology: Fiona Dick, Gonzalo S. Nido, Gry Hilde Nilsen, Christian Dölle.

Project administration: Charalampos Tzoulis.

Resources: Charalampos Tzoulis.

Software: Fiona Dick.

Supervision: Charalampos Tzoulis.

Visualization: Fiona Dick, Christian Dölle.

Writing – original draft: Fiona Dick.

Writing – review & editing: Fiona Dick, Gonzalo S. Nido, Guido Werner Alves, Ole-Bjørn Tysnes, Christian Dölle, Charalampos Tzoulis.

References

1. Tysnes OB, Storstein A. Epidemiology of Parkinson's disease. Journal of Neural Transmission. 2017; 124(8):901–905. https://doi.org/10.1007/s00702-017-1686-y PMID: 28150045

2. Borrageiro G, Haylett W, Seedat S, Kuivaniemi H, Bardien S. A review of genome-wide transcriptomics studies in Parkinson's disease. European Journal of Neuroscience. 2018; 47(1):1–16. PMID: 29068110
3. Elkon R, Ugalde AP, Agami R. Alternative cleavage and polyadenylation: extent, regulation and function. Nature Reviews Genetics. 2013; 14(7):496. https://doi.org/10.1038/nrg3482 PMID: 23774734
4. Gruber AJ, Zavolan M. Alternative cleavage and polyadenylation in health and disease. Nature Reviews Genetics. 2019; p. 1. PMID: 31267064
5. Reyes A, Huber W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. Nucleic acids research. 2017; 46(2):582–592. https://doi.org/10.1093/nar/gkx1165
6. Soneson C, Matthes KL, Nowicka M, Law CW, Robinson MD. Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. Genome biology. 2016; 17(1):12. https://doi.org/10.1186/s13059-015-0862-3 PMID: 26813113
7. Hefti MM, Farrell K, Kim S, Bowles KR, Fowkes ME, Raj T, et al. High-resolution temporal and regional mapping of MAPT expression and splicing in human brain development. PloS one. 2018; 13(4):e0195771. https://doi.org/10.1371/journal.pone.0195771 PMID: 29634760
8. Vitting-Seerup K, Sandelin A. The landscape of isoform switches in human cancers. Molecular Cancer Research. 2017; 15(9):1206–1220. https://doi.org/10.1158/1541-7786.MCR-16-0459 PMID: 28584021
9. Lin L, Park JW, Ramachandran S, Zhang Y, Tseng YT, Shen S, et al. Transcriptome sequencing reveals aberrant alternative splicing in Huntington's disease. Human molecular genetics. 2016; 25(16):3454–3466. https://doi.org/10.1093/hmg/ddw187 PMID: 27378699
10. Rhinn H, Qiang L, Yamashita T, Rhee D, Zolin A, Vanti W, et al. Alternative α-synuclein transcript usage as a convergent mechanism in Parkinson's disease pathology. Nature communications. 2012; 3:1084. https://doi.org/10.1038/ncomms2032 PMID: 23011138
11. La Cognata V, D'Agata V, Cavalcanti F, Cavallaro S. Splicing: is there an alternative contribution to Parkinson's disease? Neurogenetics. 2015; 16(4):245–263. https://doi.org/10.1007/s10048-015-0449-x PMID: 25980689
12. Beyer K, Domingo-Sàbat M, Humbert J, Carrato C, Ferrer I, Ariza A. Differential expression of alpha-synuclein, parkin, and synphilin-1 isoforms in Lewy body disease. Neurogenetics. 2008; 9(3):163–172. https://doi.org/10.1007/s10048-008-0124-6 PMID: 18335262
13. Humbert J, Beyer K, Carrato C, Mate JL, Ferrer I, Ariza A. Parkin and synphilin-1 isoform expression changes in Lewy body diseases. Neurobiology of disease. 2007; 26(3):681–687. https://doi.org/10.1016/j.nbd.2007.03.007 PMID: 17467279
14. Lin X, Cook TJ, Zabetian CP, Leverenz JB, Peskind ER, Hu SC, et al. DJ-1 isoforms in whole blood as potential biomarkers of Parkinson disease. Scientific reports. 2012; 2:954. https://doi.org/10.1038/srep00954 PMID: 23233873
15. Alves G, Müller B, Herlofson K, HogenEsch I, Telstad W, Aarsland D, et al. Incidence of Parkinson's disease in Norway: the Norwegian ParkWest study. Journal of Neurology, Neurosurgery & Psychiatry. 2009; 80(8):851–857. https://doi.org/10.1136/jnnp.2008.168211
16. Nowicka M, Robinson MD. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Research. 2016; 5. https://doi.org/10.12688/f1000research.8900.1 PMID: 28105305
17. Reyes A, Anders S, Huber W. Inferring differential exon usage in RNA-Seq data with the DEXSeq package; 2013.
18. Nido GS, Dick F, Toker L, Petersen K, Alves G, Tysnes OB, et al. Common gene expression signatures in Parkinson's disease are driven by changes in cell composition. Acta Neuropathologica Communications. 2020; 8(1):55. https://doi.org/10.1186/s40478-020-00932-7 PMID: 32317022
19. Zhuravleva E, Gut H, Hynx D, Marcellin D, Bleck CK, Genoud C, et al. Acyl coenzyme A thioesterase Them5/Acot15 is involved in cardiolipin remodeling and fatty liver development. Molecular and cellular biology. 2012; 32(14):2685–2697. https://doi.org/10.1128/MCB.00312-12 PMID: 22586271
20. Paradies G, Paradies V, De Benedictis V, Ruggiero FM, Petrosillo G. Functional role of cardiolipin in mitochondrial bioenergetics. Biochimica et Biophysica Acta (BBA)-Bioenergetics. 2014; 1837(4):408–417. https://doi.org/10.1016/j.bbabio.2013.10.006 PMID: 24183692
21. Burté F, Houghton D, Lowes H, Pyle A, Nesbitt S, Yarnall A, et al. Metabolic profiling of Parkinson's disease and mild cognitive impairment. Movement Disorders. 2017; 32(6):927–932. PMID: 28394042
22. Kaji S, Maki T, Kinoshita H, Uemura N, Ayaki T, Kawamoto Y, et al. Pathological endogenous α-synuclein accumulation in oligodendrocyte precursor cells potentially induces inclusions in multiple system

atrophy. Stem cell reports. 2018; 10(2):356–365. https://doi.org/10.1016/j.stemcr.2017.12.001 PMID: 29337114

23. Lee Y, Morrison BM, Li Y, Lengacher S, Farah MH, Hoffman PN, et al. Oligodendroglia metabolically support axons and contribute to neurodegeneration. Nature. 2012; 487(7408):443. https://doi.org/10.1038/nature11314 PMID: 22801498

24. Ramanan VK, Risacher SL, Nho K, Kim S, Swaminathan S, Shen L, et al. APOE and BCHE as modulators of cerebral amyloid deposition: a florbetapir PET genome-wide association study. Molecular psychiatry. 2014; 19(3):351–357. https://doi.org/10.1038/mp.2013.19 PMID: 23419831

25. Lockridge O, Masson P. Pesticides and susceptible populations: people with butyrylcholinesterase genetic variants may be at risk. Neurotoxicology. 2000; 21(1-2):113–126. PMID: 10794391

26. Rösler TW, Salama M, Shalash AS, Khedr EM, El-Tantawy A, Fawi G, et al. K-variant BCHE and pesticide exposure: Gene-environment interactions in a case–control study of Parkinson's disease in Egypt. Scientific reports. 2018; 8(1):16525. https://doi.org/10.1038/s41598-018-35003-4

27. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

28. Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. FastQC; 2012. Babraham Institute.

29. Patro R, Duggal G, Kingsford C. Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment. BioRxiv. 2015; p. 021592.

30. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015; 4. https://doi.org/10.12688/f1000research.7563.1 PMID: 26925227

31. Love MI, Soneson C, Patro R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Research. 2018; 7. https://doi.org/10.12688/f1000research.15398.1 PMID: 30356428

32. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. Genome research. 2012; 22(10):2008–2017. https://doi.org/10.1101/gr.133744.111 PMID: 22722343

33. Mancarci BO, Toker L, Tripathy SJ, Li B, Rocco B, Sibille E, et al. Cross-laboratory analysis of brain cell type transcriptomes with applications to interpretation of bulk tissue data. Eneuro. 2017; 4(6). https://doi.org/10.1523/ENEURO.0212-17.2017 PMID: 29204516

34. Toker L, Mancarci BO, Tripathy S, Pavlidis P. Transcriptomic evidence for alterations in astrocytes and parvalbumin interneurons in subjects with bipolar disorder and schizophrenia. Biological psychiatry. 2018; 84(11):787–796. https://doi.org/10.1016/j.biopsych.2018.07.010 PMID: 30177255

35. Van den Berge K, Soneson C, Robinson MD, Clement L. stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage. Genome biology. 2017; 18(1):151. https://doi.org/10.1186/s13059-017-1277-0 PMID: 28784146

36. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research. 2019; 47(D1):D607–D613. https://doi.org/10.1093/nar/gky1131 PMID: 30476243

37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nature genetics. 2000; 25(1):25–29. https://doi.org/10.1038/75556 PMID: 10802651

38. Consortium GO. The gene ontology resource: 20 years and still GOing strong. Nucleic acids research. 2019; 47(D1):D330–D338. https://doi.org/10.1093/nar/gky1055

39. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017; 33(21):3387–3395. https://doi.org/10.1093/bioinformatics/btx431 PMID: 29036616

Paper III

Altered transcriptome-proteome coupling indicates aberrant proteostasis in Parkinson's disease

Fiona Dick, Ole-Bjørn Tysnes, Guido Werner Alves, Gonzalo S. Nido, Charalampos Tzoulis

Manuscript

Altered transcriptome-proteome coupling indicates aberrant proteostasis in Parkinson's disease

Fiona Dick[1,2], Ole-Bjørn Tysnes[1,2], Guido Werner Alves[3,4], Gonzalo S. Nido[1,2], Charalampos Tzoulis[1,2*]

1 Neuro-SysMed, Department of Neurology, Haukeland University Hospital, Bergen, Norway
2 Department of Clinical Medicine, University of Bergen, Bergen, Norway
3 The Norwegian Center for Movement Disorders and Department of Neurology, Stavanger University Hospital, Stavanger, Norway
4 Department of Mathematics and Natural Sciences, University of Stavanger, Stavanger, Norway

* charalampos.tzoulis@uib.no

Abstract

The correlation between mRNA and protein levels has been shown to decline in the ageing brain, possibly reflecting age-dependent changes in the proteostasis. It is thought that impaired proteostasis may be implicated in the pathogenesis of Parkinson's disease (PD), but evidence derived from the patient brain is currently limited. Here, we hypothesized that if impaired proteostasis occurs in PD, this should be reflected in the form of altered correlation between transcriptome and proteome compared to healthy ageing.

To test this hypothesis, we integrated transcriptomic data with proteomics from prefrontal cortex tissue of 17 PD patients and 11 demographically matched healthy controls and assessed gene-specific correlations between RNA and protein level. To control for the effects of ageing, brain samples from 4 infants were included in the analyses.

In the healthy aged brain, we observed a genome-wide decreased correlation between mRNA and protein levels. Moreover, a group of genes encoding synaptic vesicle proteins exhibited inverse correlations. This phenomenon likely reflects the spatial separation of mRNA and protein into the neuronal soma and synapsis, respectively, commonly characterizing these genes. Most genes showed a significantly lower correlation between mRNA and protein levels in PD compared to neurologically healthy ageing, consistent with a proteome-wide decline in proteostasis. Genes showing an inverse correlation in PD were enriched for proteasome subunits, suggesting that these proteins show accentuated spatial separation of transcript and protein between the soma and axon/synapses in PD neurons. Moreover, the PD brain was characterized by increased positive mRNA-protein correlation for some genes encoding components of the mitochondrial respiratory chain, suggesting these may require tighter regulation in the face of mitochondrial pathology characterizing the PD brain.

Our results are highly consistent with a proteome-wide impairment of proteostasis in the PD brain and strongly support the hypothesis that aberrant proteasomal function is implicated in the pathogenesis of PD. Moreover, our findings have important implications for the correct interpretation of differential gene expression studies in PD. In the presence of disease-specific altered coupling of transcriptome and proteome, measured differences in mRNA levels cannot be used to infer changes at the protein-level and should be supplemented with direct determination of proteins nominated by the analyses.

Introduction

Gene expression is the process by which the information encoded in the genome (i.e., the genotype) determines the phenotype. Typically, information encoded in the DNA is first transcribed to RNA and then translated into protein, the functional product influencing the phenotype [42]. Despite the hierarchical organization of gene expression, the relationship between transcript and protein levels is highly variable in mammalian cells, both across genes and across individuals. Imperfect correlations between mRNA and protein levels are commonly attributed to regulatory mechanisms acting downstream of transcription and influencing the rate of protein synthesis and degradation [5, 10, 28]. For example, splicing, polyadenylation, and RNA-binding factors regulate translation rates, while the ubiquitin-proteasome system and lysosomal degradation regulate protein turnover. The balanced interplay between these regulatory mechanisms is crucial for maintaining cellular proteostasis.

It was recently shown that the correlation between mRNA and protein levels declines with ageing in the human brain, possibly due to altered post-transcriptional regulation [8, 54] and declining proteostasis [28]. Impaired proteostasis is thought to contribute to the misfolding and aggregation of proteins observed in neurons and other postmitotic cells with ageing [28], a phenomenon that is substantially more pronounced in age-dependent neurodegenerative proteinopathies, such as Parkinson's disease (PD) and Alzheimer's disease (AD) [17]. Several lines of evidence indicate that aberrant proteostasis is indeed implicated in PD [36]. The accumulation of Lewy bodies and neurites, intraneuronal inclusions containing aggregated forms of the protein α-synuclein [27], suggests decreased function of the autophagy–lysosomal pathway [37]. This is further supported by the fact that mutations in *GBA*, encoding the lysosomal enzyme glucocerebrosidase, greatly increase the risk of PD [1]. Altered mRNA levels of proteasomal components have been consistently found in transcriptomic studies of the PD brain [7], suggesting that dysfunction of the ubiquitin-proteasome system may also play a role.

We hypothesized that if impaired proteostasis occurs in PD, this should be reflected in the form of altered correlation between the transcriptome and proteome in the patients' brain compared to healthy ageing. To test our hypothesis, we performed transcriptome and proteome-wide analyses, using RNA sequencing and proteomics, in the brain of 17 PD patients and 11 demographically matched healthy controls, and assessed the correlation between the levels of each transcript and its cognate protein. Since it is known that extensive changes leading to mRNA-protein decoupling occur with ageing in the human brain [8, 54], we also analyzed brain samples of four individuals in early infancy. Ageing remains the strongest known PD risk factor, and this additional group allowed us to distinguish changes in mRNA-protein correlations arising due to neurologically healthy ageing from those that are specific to pathological ageing with PD.

Our results show that the PD brain is characterized by genome-wide altered mRNA-protein correlation, compared to neurologically healthy ageing. The pattern of this altered relationship between transcriptome and proteome is highly consistent with a disease-related impairment in proteostasis.

Materials and Methods

Cohorts

All experiments were conducted in fresh-frozen prefrontal cortex (Brodmann area 9) tissue from a total of 33 individuals comprising young infants (YG, $N = 4$, age 0-0.38 years), neurologically healthy aged individuals (HA, $N = 11$, age 63-88 years) and individuals with idiopathic Parkinson's disease (PD) ($N = 17$, age 69-95 years) from the Park-West study, a prospective population-based cohort which has been described in detail [2]. Whole-exome sequencing had been performed on all PD patients and known causes of Mendelian PD and other monogenic neurological disorders had been excluded [19]. Controls had no known neurological disease

and were matched for age and gender. Individuals with PD fulfilled the National Institute of Neurological Disorders and Stroke [20] and the UK Parkinson's disease Society Brain Bank [53] diagnostic criteria for the disease at their final visit. Ethical permission for these studies was obtained from our regional ethics committee (REK 2017/2082, 2010/1700, 131.04). Cohort demographics are listed in S1 File.

Sample collection

Brains were collected at autopsy and split sagittally along the corpus callosum. One hemisphere was fixed whole in formaldehyde and the other coronally sectioned and snap-frozen in liquid nitrogen. All samples were collected using a standard technique and fixation time of ~ 2 weeks. Subject demographics and tissue availability are provided in S1 Figure. Routine neuropathological examination including immunohistochemistry for α-synuclein, tau and beta-amyloid was performed on PD and HA brains. All PD cases showed neuropathological changes consistent with PD including degeneration of the dopaminergic neurons of the substantia nigra pars compacta in the presence of Lewy pathology. Controls had no pathological evidence of neurodegeneration.

RNA sequencing

Total RNA was extracted from prefrontal cortex tissue homogenate for all samples using RNeasy plus mini kit (Qiagen) with on-column DNase treatment according to the manufacturer's protocol. The final elution was made in 65μl of dH2O. The concentration and integrity of the total RNA were estimated by Ribogreen assay (Thermo Fisher Scientific), and Fragment Analyzer (Advanced Analytical), respectively and 500ng of total RNA was used for downstream RNA-seq applications. First, nuclear and mitochondrial rRNA was removed using Ribo-Zero™ Gold (Epidemiology) kit (Illumina, San Diego, CA) using the manufacturer's recommended protocol. Immediately after rRNA removal, RNA was fragmented and primed for the first strand synthesis using the NEBNext First Strand synthesis module (New England BioLabs Inc., Ipswich, MA). Directional second strand synthesis was performed using NEBNExt Ultra Directional second strand synthesis kit. Following this, the samples were taken into standard library preparation protocol using NEBNext DNA Library Prep Master Mix Set for Illumina with slight modifications. Briefly, end-repair was done followed by poly(A) addition and custom adapter ligation. Post-ligated materials were individually barcoded with unique in-house Genomic Services Lab (GSL) primers and amplified through 12cycles of PCR. Library quantity was assessed by Picogreen Assay (Thermo Fisher Scientific), and the library quality was estimated by utilizing a DNA High Sense chip on a Caliper Gx (Perkin Elmer). Accurate quantification of the final libraries for sequencing applications was determined using the qPCR-based KAPA Biosystems Library Quantification kit (Kapa Biosystems, Inc.). Each library was diluted to a final concentration of 12.5nM and pooled equimolar prior to clustering. One hundred twenty-five bp Paired-End (PE) sequencing was performed on an Illumina HiSeq2500 sequencer (Illumina, Inc.). RNA quality, measured by the DV200 score, varied across samples ($median_{YG} = 92$, $median_{CT} = 88$, $median_{PD} = 89$), although the difference between groups was not statistically significant ($p_{YG,CT} = 0.06$, $p_{CT,PD} = 0.74$, $p_{YG,PD} = 0.07$, Wilcoxon rank sum test).

RNA-Seq quality control and transcript abundance estimation

FASTQ files were trimmed using Trimmomatic version 0.39 [6] to remove potential Illumina adapters and low quality bases with the following parameters: `ILLUMINACLIP:truseq.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15`. FASTQ files were assessed using fastQC version 0.11.5 [3] prior to and following trimming. We used Salmon version 1.3.0 [41] to quantify the abundance at the transcript level with the fragment-level GC bias correction option (`--gcBias`) using the GENCODE Release 32 (GRCh38.p13) reference transcriptome and the GRCh38

reference genome, included as decoy [47]. Transcript counts were collapsed to gene-level using R package tximport version 1.14.2 with default parameters (i.e., countsFromAbundances = FALSE) and the GENCODE Release 32 (GRCh38.p13) annotation. Henceforth, we use the notion of *transcript* in a gene-centric sense, i.e., as the entity defined by all transcript isoforms mapped to the same gene. mtDNA-encoded genes were removed from the analysis. Genes were further filtered out if unusually highly expressed (i.e., if they accounted for more than 1% of a sample's library size in more than 50% of all the neurologically healthy samples (i.e., YG, HA)). We calculated log_2 transformed counts per million (CPM) for the pre-filtered set of genes. Low-expressed genes ($log_2 - CPM < 0.1$, in at least 80% of the samples) were also filtered out. The pre-filtered transcriptomic dataset resulted in a total of $N = 29,601$ genes. The dataset corresponding to the PD samples, added subsequently in the analyses, was filtered independently following the same filtering approaches and resulting in a total of $N = 29,363$ genes.

Lysis and protein digestion

$10\mu L$ of lysis buffer (4% SDS, $0.01M$ TRIS pH 7.6) was added to $1mg$ of brain tissue. The tissue was mechanically lysed using Precellys CK 14 ceramic beads, together with the Precellys Evolution (Bertin Corp, Rockville MD, USA). Lysed tissue was transferred to Eppendorf tubes and heated to 95°C for 5 minutes, before centrifugation at $10.000g$ for 5 minutes. The clarified supernatant was transferred to new Eppendorf tubes. Protein measurement was performed using the Pierce BCA protein assay kit (Thermo Fisher).
The samples were mixed with up to $50\mu L$ of the clarified lysate with $200\mu L$ of 8 M urea in 0.1 M Tris/HCl pH 8.5 in the filter unit (Microcon YM-30 (Millipore, Cat. MRCF0R030)) and centrifuged at 14,000 × g for 30 min and repeated twice. In total $30\mu g$ of protein per sample was used. The samples were reduced with $10mM$ DTT (1h, RT) and alkylated using $50mM$ IAA (1h, RT), and digested overnight at 37°C with 1:50 enzyme: substrate ratio of sequencing grade trypsin (Promega, Madison, WI). Following digestion, samples were acidified with formic acid and desalted using HLB Oasis SPE cartridges (Waters, Milford, MA). Samples were eluted with 80% acetonitrile in 0.1% formic acid and lyophilized. Peptides were stored at −80°C until use [26].

TMT labeling and fractionation

Digested peptides from each sample were chemically labelled with TMT reagents 10 plex (Thermo Fisher). Peptides were resuspended in a $30\mu L$ resuspension buffer containing $0.1M$ TEAB (Triethylammonium bicarbonate). TMT reagents ($0.1mg$) were dissolved in $41\mu L$ of anhydrous ACN of which $20\mu L$ was added to the peptides. Following incubation at RT for 1 h, the reaction was quenched using 5% hydroxylamine in HEPES buffer for 15 min at RT. The TMT-labeled samples were pooled at equal protein ratios followed by vacuum centrifuge to near dryness and desalting using Oasis PRIME HLB cartridges. Peptides were fractionated into 8 fractions using the Pierce High pH Reverse-phase Peptide fractionation kit (Thermo Fisher Scientific). The TMT experiment batch setup included additional samples which were not considered in the analysis but included in the preprocessing (filtering and normalization) of the proteomics data.

Liquid Chromatography and Mass Spectrometry Analysis

Each sample was freeze-dried in a Centrivap Concentrator (Labconco) and dissolved in 2% ACN, 1% FA. Approximately 0.5 μg of peptides from each fraction was injected into an Ultimate 3000 RSLC system (Thermo Scientific) connected to a Q-Exactive HF equipped with an EASY-spray ion source (Thermo Scientific). The samples were loaded and desalted on a precolumn (Acclaim PepMap 100, 2 cm ·$75\mu m$ i.d. nanoViper column, packed with 3 μm C18 beads) at a flow rate of $3\frac{\mu L}{min}$ for 5 min with 0.1% TFA. The peptides were separated during a biphasic ACN gradient

from two nanoflow UPLC pumps (flow rate of $0.200\frac{\mu L}{min}$) on a $50cm$ analytical column (PepMap RSLC, 50 cm $\cdot 75\mu m$ i.d. EASY-spray column, packed with $2\mu m$ C18 beads (Thermo Scientific). Solvent A was 0.1% FA in water, and Solvent B was 100% ACN. The mass spectrometer was operated in data-dependent acquisition mode to automatically switch between full scan MS1 and MS2 acquisition. The instrument was controlled through Q Excative HF Tune 2.4 and Xcalibur 3.0. MS spectra were acquired in the scan range of $375 - 1500$ m/z with resolution of $60,000$ at m/z 200, automatic gain control (AGC) target of $3 \cdot 10^6$, and a maximum injection time (IT) of $50ms$. The 12 most intense eluting peptides above intensity threshold $6 \cdot 10^4$, and charge states two or higher, were sequentially isolated for higher energy collision dissociation (HCD) fragmentation and MS2 acquisition to a normalized HCD collision energy of 32%, target AGC value of $1 \cdot 10^5$, resolution $R = 60,000$, and IT of 110 ms. The precursor isolation window was set to $1.6m/z$ with an isolation offset of 0.3 and a dynamic exclusion of $30s$. Lock-mass (445.12003 m/z) internal calibration was used, and isotope exclusion was active.
Raw data were analyzed by MaxQuant v1.5.5.1 [13] with "Variable Modifications" set for TMT 10-plex 126, 127N, 127C, 128N 128C, 129N, 129C, 130N, 130C, 131 to be at N-termini, as well as lysine for database searching and peptide identification.

Proteomics normalization and filtering

Aggregated protein intensities from maxQuant were further processed in a downstream analysis using R. First, proteins labelled as "Reverse", "Potential. contaminant" and "Only.identified.by.site" were removed from the analysis. In addition, proteins were removed if they exhibited at least one zero intensity in a sample. In order to filter out highly-expressed proteins, we selected the top four highest expressed proteins in each sample (which ranged from 3% to 5% of the total expression of a sample). The union set of these (a total of 19 proteins) was then filtered out from every sample.
We considered three possible normalization approaches for protein quantification, i) raw protein intensities, ii) quantile normalization, and iii) batch effect correction [9] followed by root mean square scaling. To assess each of these strategies we explored the association of the first two components of the principal component analysis (PCA) of the protein expression matrix with the TMT batch. Raw protein intensities (i) showed a clear clustering of samples which was associated with the batches of the TMT experiment, which was further amplified by quantile normalization (ii). This effect was no longer noticeable when we applied batch correction (iii), as suggested in [9], where we divided protein intensities by the correction factor based on the reference channels in the respective batches, followed by root mean square scaling (S1 Figure). Additionally, we were able to leverage the RNA-seq data from the same samples to gain insight into the biological validity of the three alternative normalization options by studying the transcriptome-proteome correlation in the neurologically healthy groups (HA and YG; log_2 transformed values for proteins, and log_2 transcript CPMs). The transcriptome-proteome correlation was significantly higher in the batch-corrected strategy both across samples and across genes (S2 Figure). Based on these observations we chose to apply the batch correction and subsequent root mean square scaling (iii). The pre-filtered proteomic dataset was composed of a total of $N = 2,961$ proteins.

Covariance between omic layers

We used sparse partial least square (sPLS) as implemented in the mixOmics R package version 6.10.9 [34,43] to find the linear combinations of variables (transcripts and proteins) that maximize covariance between the transcriptomic and the proteomic layers. sPLS was performed on the pre-filtered transcriptomic (X) and proteomic (Y) datasets using the "canonical" mode and the parameters keepX = 50 and keepY = 50 for feature selection.

Correlation between transcriptome and proteome

To investigate changes in the transcriptome-proteome correlation between neurologically-healthy groups (YG *vs* HA) we performed an additional filtering step on both transcripts and proteins, aiming at increasing the biological signal-to-noise ratio. Genes were flagged for removal if they satisfied at least one of the following criteria: i) not present in the pre-filtered transcriptome, ii) not present in the pre-filtered proteome, iii) low median transcript expression (below 10% quantile), iv) low transcript variance (below 15% quantile). The removal of flagged genes resulted in an analysis-ready dataset of $N = 2,107$ genes.
The dataset corresponding to the PD samples, employed in a subsequent comparison, was filtered independently following the same filtering approaches and resulting in a slightly lower number of genes in the final analysis-ready list ($N = 1,942$). Gene-wise transcript-protein Pearson correlations were calculated across samples independently for each group (CT, PD, YG) using log_2 transformed CPMs for transcript abundance and log2 transformed batch-corrected and root mean square scaled protein intensities.

Gene scoring

To investigate changes in the transcriptome-proteome correlation between groups, we applied different gene scoring strategies to rank genes according to their change in correlation (δr). For example, to investigate changes occurring in the healthy ageing process (i.e., comparing YG vs HA) each gene would be scored by $\delta r = r_{HA} - r_{YG}$. Correspondingly, to investigate changes occurring in the process of ageing with Parkinson's disease, gene scores would be calculated as $\delta r = r_{PD} - r_{YG}$. Finally, changes in transcript-protein correlations between CT and PD groups would be calculated as $\delta r = r_{PD} - r_{HA}$ (Figure 1A).
Specifically, for each of these three group comparisons (YG→HA, YG→PD, HA→PD), we wanted to identify genes belonging to three functional scenarios in regard to their transcript-protein coupling: a) *"decoupling"*, genes that show a positive transcript-protein correlation in the reference group (e.g., YG) and loose this correlation ($r \sim 0$) in the other group (e.g., HA); b) *"increased inverse correlation"*, genes which show a correlation above or equal to zero in the reference group and a negative correlation in the other group; and c) *"increased positive correlation"*, genes with a correlation above or equal to zero in the reference group that show an increased correlation in the group compared (Figure 1B). To this end, gene-specific scores were calculated as follows:

$$\forall R_{ref} > 0 \tag{1}$$

$$S_a^i = -|R_{ageing}| + R_{ref} \tag{2}$$

$$S_b^i = -R_{ageing} + t(R_{ref}) \tag{3}$$

$$S_c^i = R_{ageing} - t(R_{ref}), \tag{4}$$

$$with\ t(x) = \frac{x+1}{2}, \tag{5}$$

where $i \in 1, 2, 3$ specified the comparison being made:

$$R_{ref} = \begin{cases} R_{YG}, & \text{for i} \in 1, 2 \\ R_{HA}, & \text{for i} = 3 \end{cases} \tag{6}$$

$$R_{ageing} = \begin{cases} R_{PD}, & \text{for i} \in 2, 3 \\ R_{HA}, & \text{for i} = 2 \end{cases} \tag{7}$$

Heatmaps to visualize scoring distributions in Figure 1C were created with the R package ComplexHeatmap [23].

Pathway enrichment analysis and clustering for visualization

The above gene scorings were used to test for functional enrichment. To this end, we employed the gene score resampling method implemented in the R package ermineR version 1.0.1.9, an R wrapper package for ermineJ [35] with the complete Gene Ontology (GO) database annotation [4] (using aspects: biological process, molecular function and cellular component).

Protein interaction networks

Protein-protein interaction networks were generated using the R package coexnet version 1.8.0 [25], which retrieves information on protein co-expression and experimentally evidenced interaction from STRING [50]. Vertices were clustered using the R package igraph version 1.2.5 [14], and its implemented "edge-betweenness" cluster algorithm.

Results

Brain RNA and protein expression patterns are highly distinct between neurodevelopment and healthy ageing

Using RNA-seq and LC-MS-based proteomics, we mapped the transcriptome and proteome in prefrontal cortex tissue from 4 young infants (YG), 11 neurologically healthy aged individuals (HA), and 17 individuals with idiopathic PD (PD). First, we assessed the overall expression pattern of the groups YG and HA by integrating gene expression (X, $N = 29,601$ genes) with protein expression (Y, $N = 2,918$ proteins). Using sparse Partial Least Squares regression (sPLS), we were able to reduce dimensionality for both X and Y and project the samples in an unsupervised manner onto the combined XY-variate space. The groups YG and HA were markedly separated according to their biological characteristics in the combined variate space (YG cluster median silhouette width = 0.71, Euclidean distance; HA group median silhouette width = 0.53, Euclidean distance; Figure 2A) as well as in the separated variate space (Figure 2B), meaning that the group separation was independent of whether the selected features were restricted to either the transcriptome or the proteome, with both datasets strongly agreeing. The first XY-variate, was strongly correlated with age ($r = 0.95$, $p = 8.24 \cdot 10^{-8}$, Pearson). The $N = 50$ selected features for each component (I, II) of X and Y, which were sufficient to separate the groups, are visualized in a correlation heatmap in Figure 2C, highlighting interactions between features of X and features of Y for which the correlation was greater than $r = 0.2$.

The transcriptome-proteome correlation signature is altered in the aged brain

Since mRNA and protein levels are known to be tightly correlated during neurodevelopment, we leveraged the YG group as a control outgroup to assess alterations that occur with age and/or PD. To compare the transcriptome-proteome coupling between YG and HA groups, we calculated gene-wise correlation coefficients (r, Pearson) across samples in each of the groups (r_{YG} and r_{HA}) for the YG and HA groups, respectively). After pre-filtering, we were able to assess the transcript-protein level correlation for 2,107 genes. Correlation coefficients for each group are listed in S2 File. We will henceforth use the term *gene* for both the gene and the protein it encodes.

As expected, transcriptome and proteome were significantly more correlated in the YG group compared to the HA group as shown by the transcript-protein r distributions (median $r_{YG} = 0.31$; median $r_{HA} = 0.07$; $p < 2 \cdot 10^{-16}$, Wilcoxon) (Figure 3A). To further characterize the differences in the transcriptome-proteome coupling, we generated a two-dimensional density plot of the gene-wise transcript-protein correlation in the YG and HA groups (Figure 3B). The vast majority of genes exhibited a high transcript-protein correlation in YG ($r_{YG} > 0.5$) and a lack of correlation

in HA (r_{HA}\$$sim$0). We henceforth refer to this age-dependent decrease in transcript-protein correlation as decreased coupling or, simply, *decoupling*. Additional high-density areas were observed for genes with low absolute transcript-protein correlation in both groups, and for genes transitioning from a highly positive correlation in YG to an inverse correlation in HA. Finally, very few genes showed an age-dependent increase in correlation. These observations indicate that most genes show a tight positive correlation between mRNA and protein levels during early infancy. With aging, however, this correlation either decreases towards zero ($r_{HA} \to 0$, decoupling) or becomes inverse ($r_{HA} < 0$, increased anticorrelation).

Altered mRNA-protein correlation in the aged brain is enriched in specific biological functions

Next, we assessed whether altered mRNA-protein correlation in the aged brain is enriched for specific pre-defined biological pathways. To this end, we divided genes into three groups according to their changes in correlation (YG $\to$ HA): i) decoupled ($r_{YG} > 0$, $r_{HA} \sim 0$), ii) increased inverse correlation ($r_{YG} > 0$, $r_{HA} < 0$), iii) increased positive correlation ($r_{YG} > 0$, $r_{HA} > r_{YG}$). Genes in each group were ranked according to the magnitude of the difference ($\delta(r_{HA}, r_{YG})$) (Figure 1, S2 File). While the majority of genes showed decoupling with ageing (group i), we found no significant enrichment in this group for any specific biological pathway. A protein-protein interaction network of the top decoupled genes (gene score > 90% quantile, $N = 58$), revealed 4 interconnected groups with more than 5 members (Figure 3C), strongly suggesting a functional relationship. Notably, 5 of the 7 members of one of these groups were subunits of the proteasome complex (*PSMA4*, *PSMB3*, *PSMD5*, *PSMD8*, *PSMD14*). The gene group with increased inverse correlation (group ii) showed significant enrichment for 14 pathways (FDR < 0.05), mostly related to synaptic components and synaptic vesicles (S2 Figure, sheet S1b). Finally, the minority of genes which showed increased positive correlation from YG to HA (iii) were enriched in "regulation of hemostasis" (FDR = 0.04). Significantly enriched GO terms for each of the three groups are listed with their FDR adjusted p-value in (S2 Figure, sheet S1b-c).

The age-dependent decoupling between mRNA and protein levels is more pronounced in the PD brain

Next, we wanted to assess how the coupling between transcriptome and proteome changes in PD compared to normal, neurologically healthy ageing. Transcript-protein correlations across all three groups (YG, HA and PD) were assessed for a total of 1,907 genes (see Methods). The correlation distributions for PD and HA groups showed no significant difference ($p = 0.52$, Wilcoxon) with a median close to zero for both groups (median $r_{PD} = 0.10$, median $r_{HA} = 0.08$). However, PD exhibited an overall lower variance ($\sigma^2(r_{HA}) = 0.15$, $\sigma^2(r_{PD}) = 0.08$) and a reduced range ($range(r_{HA}) = [-0.97, 0.93]$; $range(r_{PD}) = [-0.70, 0.80]$), suggesting a more pronounced trend of decoupling (Figure 4A).

To further investigate this, we calculated the absolute difference in the gene-wise transcript-protein correlation between the YG group and either the HA ($\delta_{age} = |r_{HA}| - |r_{YG}|$) or the PD group ($\delta_{PD} = |r_{PD}| - |r_{YG}|$). Interestingly, the two distributions differed significantly ($p = 2.2 \cdot 10^{-16}$, Wilcoxon, paired), with δ_{PD} being significantly higher than $delta_{age}$ (Figure 4B). These findings indicate that the age-dependent loss of transcript-protein correlation is more pronounced in pathological aging with PD than in healthy aging, as evident also by the $r_{HA} \sim r_{YG}$ and $r_{PD} \sim r_{YG}$ density distributions (S3 Figure). Despite these differences, δ_{age} and δ_{PD} showed a highly significant positive correlation ($r = 0.71$, $p < 2.2 \cdot 10^{-16}$, Pearson) (Figure 4C), suggesting that the process of decoupling is qualitative similar and has a comparable genome-wide distribution in HA and PD, although it is more pronounced in the latter.

Altered transcript-protein correlation in the PD brain is enriched for specific biological processes

Similar to healthy ageing, transcript-protein correlation in the PD-brain showed a general trend for decoupling compared to the YG group (Figure 4D) and no significant enrichment for specific pathways. Genes showing increased inverse correlation with PD ageing were significantly enriched for 136 pathways (FDR < 0.05) mainly related to protein degradation (including proteasome complex, ubiquitination and unfolded protein response), immune response, transcription and cell-cycle regulation (S2 Figure, sheet S2b). However, a closer look at the enrichment results singled out several proteasomal subunits as main drivers of the enrichment signal across all these pathways. Removal of proteasomal subunits from the gene lists, in fact, resulted in only three statistically significant pathways. The signal of two of them (related to enhancer-binding), was driven solely by 5 genes (*CALCOCO1*, *H3F3A*, *RUVBL2*, *SFPQ*, *SUB1*), while the enrichment of the third significant pathway (unfolded protein binding) was supported by 36 genes which showed negative correlation in PD. Genes showing increased positive correlation with PD ageing were significantly enriched for 46 pathways, most of which were related to mitochondrial respiration (S2 Figure, sheet S2c).

Since comparing YG and PD cannot confidently differentiate ageing- from disease-related changes, we also performed a direct comparison between HA and PD. These analyses revealed an altered profile of transcript-protein correlation in PD compared to HA (Figure 4D). Genes showing increased decoupling in PD were significantly enriched for 17 pathways mostly related to nitrogen metabolism. Genes showing increased inverse correlation in PD were enriched in 54 biological pathways, primarily related to protein degradation and immune response. Similar to the comparison with the YG-group, this enrichment was driven primarily by proteasomal subunits. The magnitude of anticorrelation was, however, heterogeneous, affecting certain subunits more than others ($median(r) = -0.18$, $range(r) = [-0.64, 0.42]$, $\sigma_r = 0.27$, $N = 22$). Finally, genes with increased positive correlation in PD were enriched for biological processes related to mitochondrial respiration. A list of significantly enriched pathways in PD compared to HA is provided in S2 Figure, sheet S3a-c.

Discussion

Here, we assess for the first time the genome-wide correlation between the transcriptome and proteome in the PD brain, compared to neurologically healthy ageing. In the infants, the vast majority of genes showed a strong positive correlation between mRNA and protein levels, suggesting that in the neonatal brain, protein abundance is determined mainly by transcript concentration. This correlation was significantly lower in the neurologically healthy aged individuals, consistent with an age-dependent decoupling between transcript and protein abundance. Similar trends have been shown in yeast [31], fish [32], and the macaque and human brain [8, 54]. Previous studies have suggested that age-dependent decoupling in the brain may preferentially affect certain biological processes, including transcriptional, translational and posttranslational regulation, signalling pathways, and mitochondrial function [31, 32, 54]. In our data, genes that decoupled in the aged group did not exhibit a significant enrichment in any specific biological pathways, suggesting that the age-dependent loss of correlation between mRNA and protein is a general, genome-wide process.

The phenomenon of age-dependent decoupling between mRNA and protein suggests that, in the ageing brain, modulating the rates of translation and protein degradation assumes a more central role in determining protein abundance than transcriptional regulation. On the other hand, the observed tight correlation between mRNA and protein levels in the neonatal brain may be, at least partly, related to the ongoing proliferation and migration of glial progenitors [49], a process heavily dependent on transcriptional regulation via the binding of a broad spectrum of transcription factors [15].

In addition to the physiological effects of brain development, the mRNA-protein decoupling observed in the ageing brain may also reflect pathological changes taking place in ageing postmitotic cells. A decline in proteasome function with ageing has been shown in multiple mammalian tissues and is believed to be contributing to the accumulation of misfolded and damaged proteins in the ageing brain (reviewed in [51]). Although not statistically significant, several subunits of the proteasomal complex were among the top decoupled during healthy ageing. Our findings provide further support to the notion of aberrant proteasomal function in the aged brain.

The age-dependent decoupling between mRNA and protein levels was significantly more pronounced in the brain of individuals with PD. While our data cannot elucidate the molecular mechanisms underlying this phenomenon, a state of heightened decoupling is consistent with disease-related impairment in proteostasis due to altered proteasomal and/or lysosomal function, both of which have been implicated in the pathogenesis of PD by numerous studies [30, 33, 55]. Thus, our findings support the hypothesis that aberrant proteostasis contributes to the pathogenesis of PD.

In the healthy aged brain we identified a group of genes exhibiting inverse correlations between transcript and protein levels. Anticorrelation between transcript and protein readouts can be explained by the highly polarized cellular architecture of neurons, which allows spatial separation between mRNA and protein [38]. While some proteins are translated locally at their resident site, others are synthesized in the soma and transported along the axon/dendrites to their target location. This leads to a steady state in which the transcript resides in the soma, whereas most of the protein is either under transport in the axon or at the synapsis [38]. In these cases, since brain tissue samples vary in relative grey/white matter content and therefore also in relative soma/axonal content [39], readouts of transcript and protein levels will be anticorrelated across-samples. Specifically, samples enriched in somas will show a high relative transcript/protein ratio, whereas samples enriched in axons will show a low relative transcript/protein ratio. In line with this hypothesis, genes showing negative correlation between transcript and protein levels in healthy ageing were significantly enriched in synaptic vesicle related pathways. Synaptic vesicle proteins were indeed shown to be preferentially translated in the cell body and undergo axonal transport to the synapses [22, 29], consistent with a spatial compartmentalization of transcripts and their protein products. We also observed that the top negatively correlated genes in healthy ageing were highly positively correlated in the infants, which may reflect a more homogenous distribution of somata and axons and/or reduced axonal transport during development, likely due to immature neuronal morphology [21, 45].

Interestingly, genes showing inverse mRNA-protein correlation in PD were not significantly enriched in synaptic function compared to healthy ageing. At least two factors may contribute to this phenomenon. First, disruption of axonal transport has been shown to occur in the PD brain (see [52] for a review). This would decrease the spatial separation between transcript and protein, thereby blunting the negative correlation across samples. Second, the PD brain, including the prefrontal cortex, is characterized by neuronal and synaptic loss and a relative increase in glial populations [39]. It is therefore conceivable that, if the anticorrelation signal originates from neurons, it may be diluted as a result of these changes in cellular composition.

Genes showing inverse mRNA-protein correlation in PD were enriched for subunits of the proteasomal complex compared to both infants and neurologically healthy aged individuals. This finding suggests that these proteins become specifically more polarized in PD, with accentuated spatial separation of transcript and protein between soma and axon. The ubiquitin-proteasome system has a crucial role in maintaining synaptic proteostasis and modulating neurotransmission and has been shown to be enriched at the synapses [12, 16, 24, 46, 48]. Moreover, studies in mice have shown that some proteasomal subunits are translated locally at the synapses, whereas others are translated in the soma and transported to the synapses [11, 22]. Furthermore, our data indicate that the spatial mRNA-protein separation is uneven across the proteasomal subunits, suggesting a potentially altered stoichiometry of the synaptic proteasome in PD neurons. The formation of an alternative proteasome complex consisting of an additional $\alpha - 4$ subunit

($PSMA7$) in place of an $\alpha-3$ ($PSMA4$) has been shown to be involved in cellular adaption to environmental stress [40]. These subunits showed a marked disparity in their correlation values in the PD brain ($r_{PSMA7} = -0.51$; $r_{PSMA4} = -0.17$).

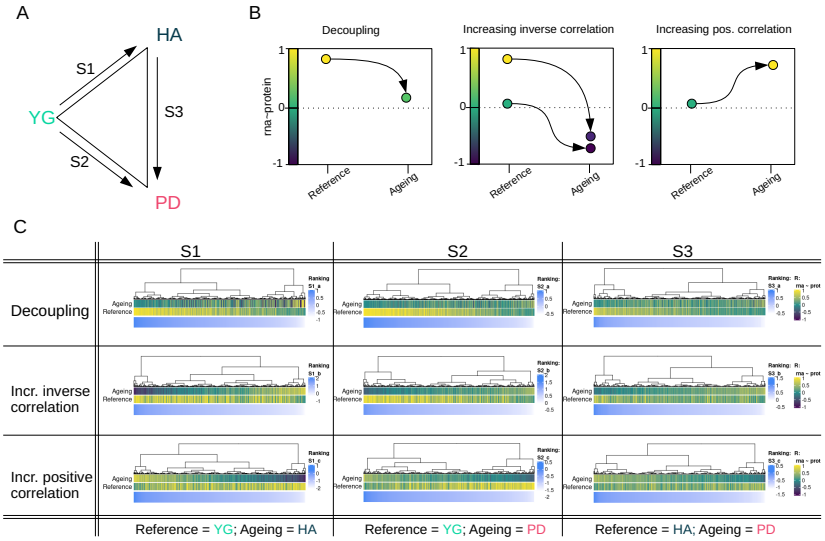
Interestingly, the PD brain was characterized by increased positive mRNA-protein correlation for genes encoding components of the mitochondrial respiratory chain. We and others have shown that quantitative and functional respiratory chain deficiencies characterize the PD-brain, including the prefrontal cortex [18, 44]. It is possible that a tighter relationship between transcription and translation of at least some of the mitochondrial respiratory chain subunits allows for better regulation in a highly strained system lacking spare capacity. Positive correlations for widely expressed genes in neurons were also observed by [38].

The findings of this study should be interpreted in light of several limitations. First, post-mortem RNA degradation in our samples may partly contribute to low correlations between mRNA and protein. Proteins are generally more resilient to post-mortem degradation and survive for longer periods than RNA. However, since there is no reason to assume that RNA degradation would be systematically different between our groups, this factor is unlikely to confound our results of differential transcript-protein correlation between groups. Second, due to the lower sensitivity of proteomics, our dataset was constrained to only 1,400 proteins. Thus, our findings are not necessarily representative of the entire genome. Third, the sample size for the YG group ($N = 4$) was small due to the limited availability of this type of tissue, limiting the generalizability of the ageing-associated findings. Nevertheless, the infant group did recapitulate the previously observed high positive correlation for the vast majority of genes [54], suggesting the samples are representative for transcript-protein correlation in the infant brain.

In summary, we show that the PD brain is characterized by altered coupling between the transcriptome and proteome, compared to neurologically healthy ageing. This altered relationship between mRNA and protein levels is consistent with an extensive, possibly proteome-wide, impairment of proteostasis, and strongly supports the hypothesis that aberrant proteasomal function is implicated in the pathogenesis of PD. Moreover, these findings have important implications for the correct interpretation of transcriptomic studies in this field. Gene expression studies are extensively used to identify disease-related pathways in ageing and neurodegeneration, and it is generally assumed that observed differences in mRNA levels reflect differences at the protein level. If the relationship between transcript and protein is altered in PD, this should be accounted for when interpreting the molecular impact of differential gene expression in the patient brain.

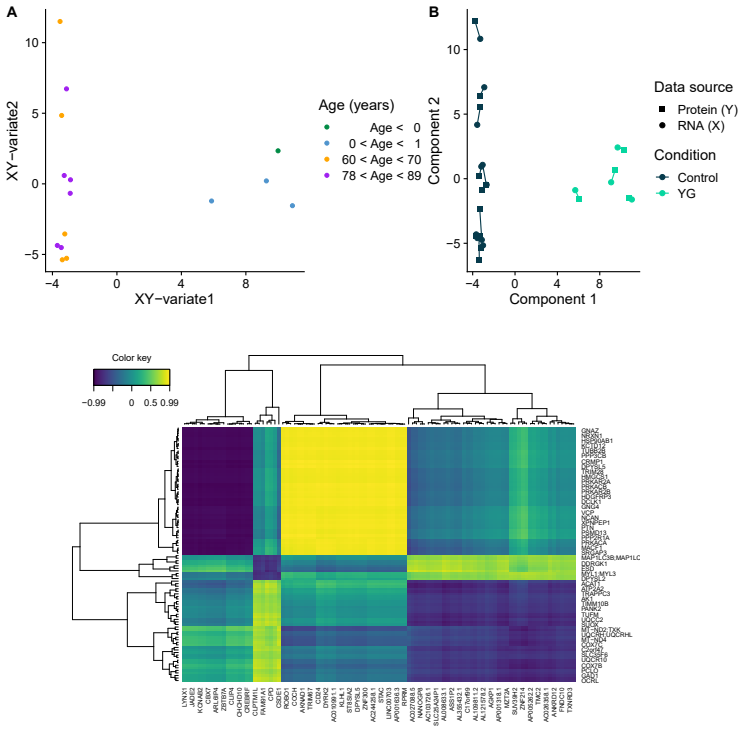
Figures

Figure 1



Gene scoring ranking for gene-set enrichment analysis
A: Schematic illustration of comparisons between groups. Each comparison is between a reference group and an ageing group (either HA or PD). For S1, we define YG as the reference and HA as the ageing group. Similarly, for S2, we define YG as a reference and PD as the ageing group. Finally, in S3 we investigate the differences between HA (reference) and PD (ageing). PD: Parkinson's disease; HA: healthy aging; YG: infants. **B:** Schematic representation of correlation changes: i) decoupling ii) increasing inverse correlation and iii) increasing positive correlation. We calculated scores to rank genes according to each of these three trends to perform change-specific pathway enrichment analysis. **C:** Gene scores calculated for the three comparisons (as defined in A) and correlation trends (as defined in B) displayed in blue, mapped to the respective reference and ageing correlation coefficient. The correlation coefficients are coloured from -1 (dark blue) to zero (green) to 1 (yellow)

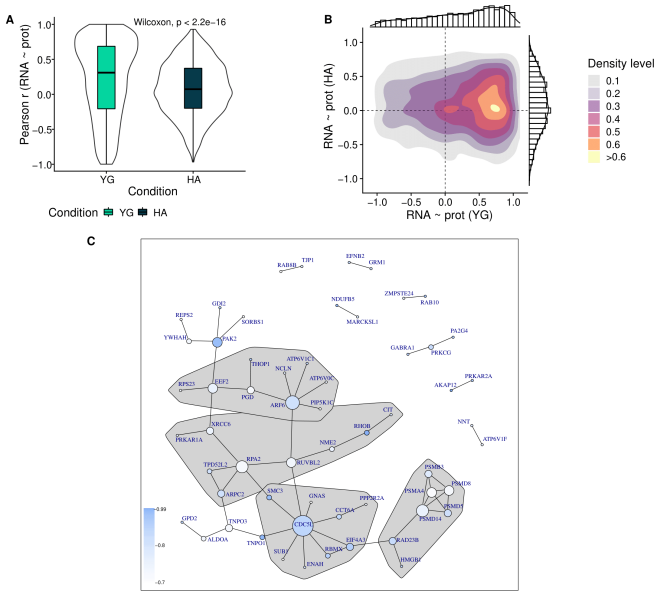
Figure 2



Integrative analysis of age-specific expression patterns in the transcriptome and proteome using sPLS

A: Data points (samples) coloured by age in years (binned) in the combined XY variate space (coordinates of samples are the mean over the coordinates in the subspaces of X and Y). **B:** Samples plotted separately in the subspaces X (circle) and Y (square) spanned by their first two components. Colour coding indicates group membership (HA: dark blue; YG: turquoise); shape indicates omic layer (protein expression: square; transcript expression: circle). **C:** Heatmap displaying the selected features of components I and II from both omic layers: transcriptome (x-axis) and proteome (y-axis). A correlation threshold of 0.2 was set to reduce the number of features in the plot and facilitate visualization. Colour indicates correlation between features of X and features of Y.

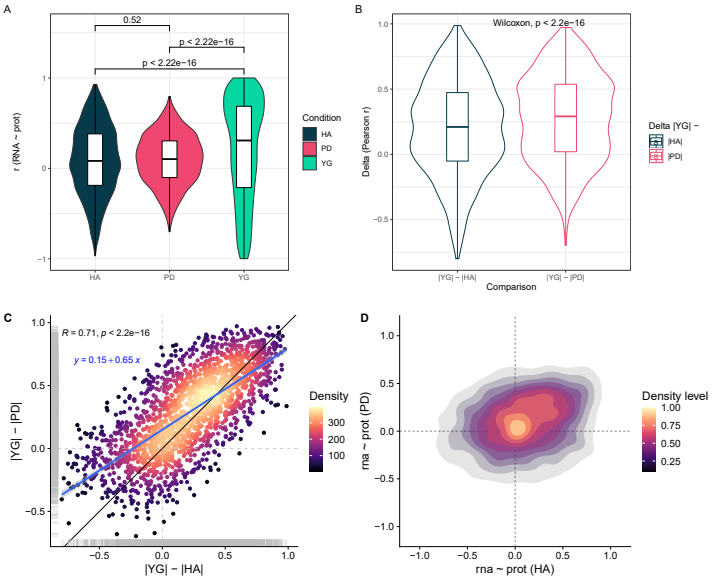
Figure 3



Decoupling of transcriptome and proteome in neurologically healthy aged individuals

A: Distribution of gene-wise correlation coefficients for the groups YG (turquoise) and HA (dark blue) (Wilcoxon unpaired test). **B:** Two-dimensional density plot displaying within-gene mRNA-protein Pearson correlations in YG (x-axis) vs HA (y-axis). **C:** Protein-protein interaction (PPI) network for genes in the 0.90 quantile of gene-scores ranking (blue) genes by decoupling in HA. Vertex communities were identified using edge betweenness (R package igraph). Only communities with more than 5 members are displayed. PPI based on coexpression, experimental evidence of interaction and neighbourhood characteristics.

Figure 4



Altered correlation coefficient distribution in PD

A: mRNA-protein correlation distributions for HA (dark blue), PD (pink) and YG (turquoise). **B:** Distribution of the deltas (differences in absolute correlation coefficients) between the reference (YG) and HA (dark blue), and YG and PD (pink). **C:** Relationship between δ_{age} (x-axis) and δ_{PD} (y-axis). Color indicates data point density. Blue line indicated the linear model fit ($y \sim x$). Black line is the diagonal (intercept = 0, slope = 1) **D:** Two-dimensional density plot displaying both distribution and relationship between the RNA $\sim$ protein correlations in HA (x-axis) and PD (y-axis).

References

1. J. Aharon-Peretz, H. Rosenbaum, and R. Gershoni-Baruch. Mutations in the glucocerebrosidase gene and parkinson's disease in ashkenazi jews. *New England Journal of Medicine*, 351(19):1972–1977, 2004.

2. G. Alves, B. Müller, K. Herlofson, I. HogenEsch, W. Telstad, D. Aarsland, O.-B. Tysnes, and J. P. Larsen. Incidence of parkinson's disease in norway: the norwegian parkwest study. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(8):851–857, 2009.

3. S. Andrews et al. Fastqc: a quality control tool for high throughput sequence data, 2010.

4. M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

5. W. E. Balch, R. I. Morimoto, A. Dillin, and J. W. Kelly. Adapting proteostasis for disease intervention. *science*, 319(5865):916–919, 2008.

6. A. M. Bolger, M. Lohse, and B. Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

7. G. Borrageiro, W. Haylett, S. Seedat, H. Kuivaniemi, and S. Bardien. A review of genome-wide transcriptomics studies in parkinson's disease. *European Journal of Neuroscience*, 47(1):1–16, 2018.

8. M. S. Breen, S. Ozcan, J. M. Ramsey, Z. Wang, A. Ma'ayan, N. Rustogi, M. G. Gottschalk, M. J. Webster, C. S. Weickert, J. D. Buxbaum, et al. Temporal proteomic profiling of postnatal human cortical development. *Translational psychiatry*, 8(1):1–14, 2018.

9. A. Brenes, J. Hukelmann, D. Bensaddek, and A. I. Lamond. Multibatch tmt reveals false positives, batch effects and missing values. *Molecular & Cellular Proteomics*, 18(10):1967–1980, 2019.

10. C. Buccitelli and M. Selbach. mrnas, proteins and the emerging principles of gene expression control. *Nature Reviews Genetics*, 21(10):630–644, 2020.

11. R. Cagnetta, C. K. Frese, T. Shigeoka, J. Krijgsveld, and C. E. Holt. Rapid cue-specific remodeling of the nascent axonal proteome. *Neuron*, 99(1):29–46, 2018.

12. I. J. Cajigas, T. Will, and E. M. Schuman. Protein homeostasis and synaptic plasticity. *The EMBO journal*, 29(16):2746–2752, 2010.

13. J. Cox and M. Mann. Maxquant enables high peptide identification rates, individualized ppb-range mass accuracies and proteome-wide protein quantification. *Nature biotechnology*, 26(12):1367–1372, 2008.

14. M. G. Csardi. Package 'igraph'. *Last accessed*, 3(09):2013, 2013.

15. E. H. Davidson. Emerging properties of animal gene regulatory networks. *Nature*, 468(7326):911–920, 2010.

16. A. Deglincerti, Y. Liu, D. Colak, U. Hengst, G. Xu, and S. R. Jaffrey. Coupled local translation and degradation regulate growth cone collapse. *Nature communications*, 6(1):1–12, 2015.

17. D. Dickson and R. O. Weller. *Neurodegeneration: the molecular pathology of dementia and movement disorders*. John Wiley & Sons, 2011.

18. I. H. Fløtnes, E. Fernandez-Vizarra, M. Lykouri, B. Brakedal, G. O. Skeie, H. Miletic, P. K. Lilleng, G. Alves, O.-B. Tysnes, K. Haugarvoll, et al. Neuronal complex i deficiency occurs throughout the parkinson's disease brain, but is not associated with neurodegeneration or mitochondrial dna damage. *Acta neuropathologica*, 135(3):409–425, 2018.

19. J. J. Gaare, G. S. Nido, P. Sztromwasser, P. M. Knappskog, O. Dahl, M. Lund-Johansen, J. Maple-Grødem, G. Alves, O.-B. Tysnes, S. Johansson, et al. Rare genetic variation in mitochondrial pathways influences the risk for parkinson's disease. *Movement Disorders*, 33(10):1591–1600, 2018.

20. D. J. Gelb, E. Oliver, and S. Gilman. Diagnostic criteria for parkinson disease. *Archives of neurology*, 56(1):33–39, 1999.

21. J. H. Gilmore, F. Shi, S. L. Woolson, R. C. Knickmeyer, S. J. Short, W. Lin, H. Zhu, R. M. Hamer, M. Styner, and D. Shen. Longitudinal development of cortical and subcortical gray matter from birth to 2 years. *Cerebral cortex*, 22(11):2478–2485, 2012.

22. C. Glock, A. Biever, G. Tushev, I. Bartnik, B. Nassim-Assir, S. T. Dieck, and E. M. Schuman. The mrna translation landscape in the synaptic neuropil. *bioRxiv*, 2020.

23. Z. Gu, R. Eils, and M. Schlesner. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, 32(18):2847–2849, 2016.

24. A. N. Hegde. Proteolysis, synaptic plasticity and memory. *Neurobiology of learning and memory*, 138:98–110, 2017.

25. J. D. Henao. coexnet: An r package to build co-expression networks from microarray data. 2018.

26. M. Hernandez-Valladares, E. Aasebø, O. Mjaavatten, M. Vaudel, Ø. Bruserud, F. Berven, and F. Selheim. Reliable fasp-based procedures for optimal quantitative proteomic and phosphoproteomic analysis on samples from acute myeloid leukemia patients. *Biological procedures online*, 18(1):13, 2016.

27. B. A. Hijaz and L. A. Volpicelli-Daley. Initiation and propagation of α-synuclein aggregation in the nervous system. *Molecular neurodegeneration*, 15(1):1–12, 2020.

28. M. S. Hipp, P. Kasturi, and F. U. Hartl. The proteostasis network and its decline in ageing. *Nature reviews Molecular cell biology*, 20(7):421–435, 2019.

29. C. E. Holt, K. C. Martin, and E. M. Schuman. Local translation in neurons: visualization and function. *Nature structural & molecular biology*, 26(7):557–566, 2019.

30. X. Hou, J. O. Watzlawik, F. C. Fiesel, and W. Springer. Autophagy in parkinson's disease. *Journal of molecular biology*, 432(8):2651–2672, 2020.

31. G. E. Janssens, A. C. Meinema, J. Gonzalez, J. C. Wolters, A. Schmidt, V. Guryev, R. Bischoff, E. C. Wit, L. M. Veenhoff, and M. Heinemann. Protein biogenesis machinery is a driver of replicative aging in yeast. *elife*, 4:e08527, 2015.

32. E. Kelmer Sacramento, J. M. Kirkpatrick, M. Mazzetto, M. Baumgart, A. Bartolome, S. Di Sanzo, C. Caterino, M. Sanguanini, N. Papaevgeniou, M. Lefaki, et al. Reduced proteasome activity in the aging brain results in ribosome stoichiometry loss and aggregation. *Molecular systems biology*, 16(6):e9596, 2020.

33. J. D. Lane, V. I. Korolchuk, J. T. Murray, C. Karabiyik, M. J. Lee, and D. C. Rubinsztein. Autophagy impairment in parkinson's disease. *Essays in biochemistry*, 61(6):711–720, 2017.

34. K.-A. Lê Cao, D. Rossouw, C. Robert-Granié, and P. Besse. A sparse pls for variable selection when integrating omics data. *Statistical applications in genetics and molecular biology*, 7(1), 2008.

35. H. K. Lee, W. Braynen, K. Keshav, and P. Pavlidis. Erminej: tool for functional analysis of gene expression data sets. *BMC bioinformatics*, 6(1):269, 2005.

36. Š. Lehtonen, T.-M. Sonninen, S. Wojciechowski, G. Goldsteins, and J. Koistinaho. Dysfunction of cellular proteostasis in parkinson's disease. *Frontiers in neuroscience*, 13:457, 2019.

37. M. A. Lynch-Day, K. Mao, K. Wang, M. Zhao, and D. J. Klionsky. The role of autophagy in parkinson's disease. *Cold Spring Harbor perspectives in medicine*, 2(4):a009357, 2012.

38. C. P. Moritz, T. Mühlhaus, S. Tenzer, T. Schulenborg, and E. Friauf. Poor transcript-protein correlation in the brain: negatively correlating gene products reveal neuronal polarity as a potential cause. *Journal of neurochemistry*, 149(5):582–604, 2019.

39. G. S. Nido, F. Dick, L. Toker, K. Petersen, G. Alves, O.-B. Tysnes, I. Jonassen, K. Haugarvoll, and C. Tzoulis. Common gene expression signatures in parkinson's disease are driven by changes in cell composition. *Acta Neuropathologica Communications*, 8(1):55, 2020.

40. A. Padmanabhan, S. A.-T. Vuong, and M. Hochstrasser. Assembly of an evolutionarily conserved alternative proteasome isoform in human cells. *Cell reports*, 14(12):2962–2974, 2016.

41. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, 14(4):417–419, 2017.

42. K. Roberts, B. Alberts, A. Johnson, P. Walter, and T. Hunt. Molecular biology of the cell. *New York: Garland Science*, 2002.

43. F. Rohart, B. Gautier, A. Singh, and K.-A. Lê Cao. mixomics: An r package for 'omics feature selection and multiple data integration. *PLoS computational biology*, 13(11):e1005752, 2017.

44. A. Schapira, J. Cooper, D. Dexter, P. Jenner, J. Clark, and C. Marsden. Mitochondrial complex i deficiency in parkinson's disease. *The Lancet*, 333(8649):1269, 1989.

45. J. C. Silbereis, S. Pochareddy, Y. Zhu, M. Li, and N. Sestan. The cellular and molecular landscapes of the developing human central nervous system. *Neuron*, 89(2):248–268, 2016.

46. S. D. Speese, N. Trotta, C. K. Rodesch, B. Aravamudan, and K. Broadie. The ubiquitin proteasome system acutely regulates presynaptic protein turnover and synaptic efficacy. *Current biology*, 13(11):899–910, 2003.

47. A. Srivastava, L. Malik, H. Sarkar, M. Zakeri, F. Almodaresi, C. Soneson, M. I. Love, C. Kingsford, and R. Patro. Alignment and mapping methodology influence transcript abundance estimation. *Genome biology*, 21(1):1–29, 2020.

48. O. Steward and E. M. Schuman. Compartmentalized synthesis and degradation of proteins in neurons. *Neuron*, 40(2):347–359, 2003.

49. J. Stiles and T. L. Jernigan. The basics of brain development. *Neuropsychology review*, 20(4):327–348, 2010.

50. D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork, et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, page gkw937, 2016.

51. V. A. Vernace, L. Arnaud, T. Schmidt-Glenewinkel, and M. E. Figueiredo-Pereira. Aging perturbs 26s proteasome assembly in drosophila melanogaster. *The FASEB Journal*, 21(11):2672–2682, 2007.

52. L. A. Volpicelli-Daley. Effects of α-synuclein on axonal transport. *Neurobiology of disease*, 105:321–327, 2017.

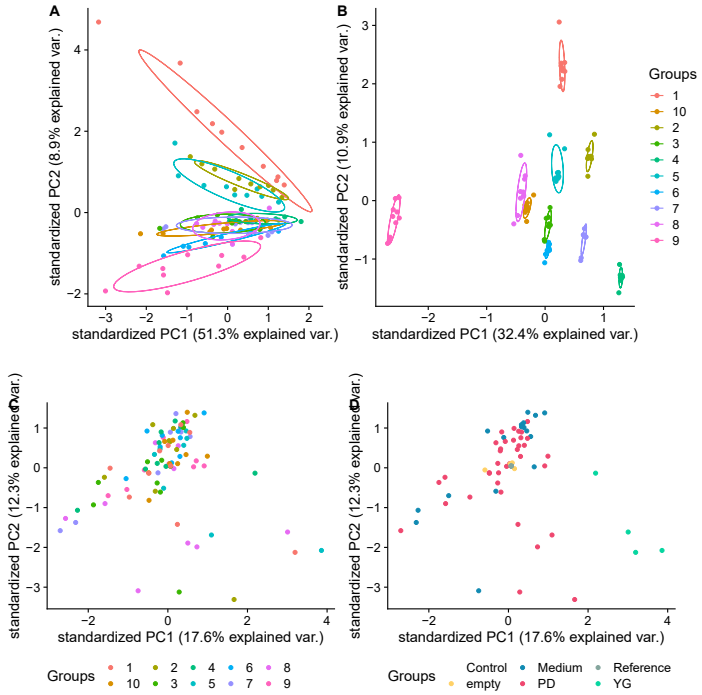
53. C. Ward. Research diagnostic criteria for parkinson's disease. *Advance in Neurology*, 53:245–249, 1990.

54. Y.-N. Wei, H.-Y. Hu, G.-C. Xie, N. Fu, Z.-B. Ning, R. Zeng, and P. Khaitovich. Transcript and protein expression decoupling reveals rna binding proteins and mirnas as potential modulators of human aging. *Genome biology*, 16(1):1–15, 2015.

55. Q. Zheng, T. Huang, L. Zhang, Y. Zhou, H. Luo, H. Xu, and X. Wang. Dysregulation of ubiquitin-proteasome system in neurodegenerative diseases. *Frontiers in aging neuroscience*, 8:303, 2016.

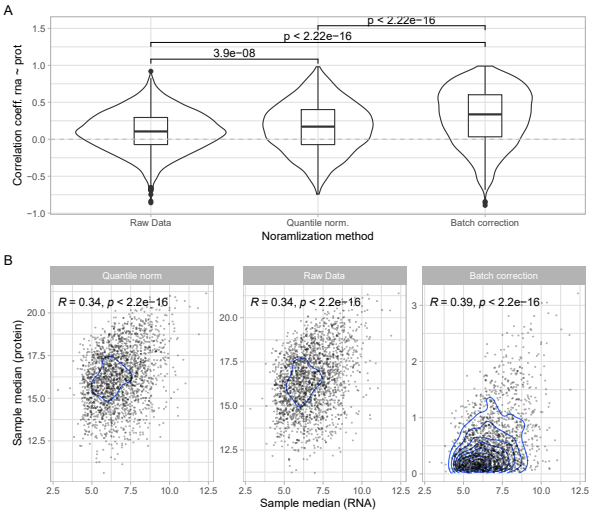
Supporting Information

S1 Figure



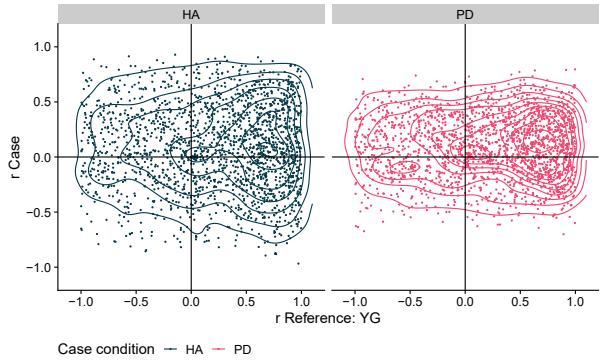
Principal component analysis visualized for all TMT samples
Data points represent samples spanned by the first (x-axis) and second (y-axis) component of principal component analysis on raw protein intensities (**A**), quantile normalized protein intensities (**B**) and scaled batch corrected protein intensities (**C** and **D**). Colouring indicates the TMT batch of the sample for A, B and C and the sample's condition for D.

S2 Figure



Batch correction of proteomics data improves correlation with RNA
A: Distribution of gene-wise correlation coefficients (RNA ∼ protein) (y-axis) are displayed for the three normalization approaches of protein intensities (x-axis). **B:** Comparison of correlation between sample-median RNA expression (x-axis) and sample-median protein expression (y-axis) for the three different protein intensity normalization approaches (facets).

S3 Figure



Distributions of RNA-protein correlations in a comparison between groups
Two-dimensional density plot displaying both distribution and relationship between the reference YG (x-axis) and the ageing groups (y-axis): YG vs HA (dark blue, first panel), and YG vs PD (pink, second panel).

S1 File

Cohort demographic and experimental information

S2 File

Correlation coefficients and gene ranking
Contained are gene-wise Pearson correlation coefficients for each group (YG, HA, PD) as well as scorings used to rank genes in the pathway enrichment analyses.

S3 File

Significantly enriched GO terms
Enriched go terms for each gene scoring are listed in respective sheets.

Acknowledgments

Mass spectrometry-based proteomic analyses were performed by (The proteomics Unit at University of Bergen (PROBE)). This facility is a member of the National Network of Advanced Proteomics Infrastructure (NAPI), which is funded by the Research Council of Norway INFRASTRUKTUR-program (project number: 295910).
We are grateful to patients and their families for participating in our research. We would also like to thank our colleagues at the Neuromics group for the fruitful discussions.

Author contributions

- Conceptualization: F. Dick, G. S. Nido, C. Tzoulis
- Methodology: F. Dick. G. S. Nido
- Software: F. Dick
- Formal Analysis: F. Dick, G. S. Nido
- Data Curation: F. Dick, G. S. Nido
- Original Draft Preparation: F. Dick
- Review and Editing: F. Dick, C. Tzoulis, G. S. Nido
- Visualization: F. Dick
- Provision of material: G. W. Alves, O. Tysnes
- Supervision: C. Tzoulis
- Project Administration: C. Tzoulis
- Funding Acquisition: C. Tzoulis

Data availability

The datasets supporting the conclusions of this article are included within the article and its supplementary files. The source code for the analyses is available on `https://github.com/fifdick/rna_protein_coupling_PD`.

Funding

This work is supported by grants from The Research Council of Norway (288164, ES633272) (https://www.forskningsradet.no/en/) and Bergen Research Foundation (BFS2017REK05) (https://mohnfoundation.no/engelsk-rekruttering/?lang=en). Both of these were received by CT. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors have declared that no competing interests exist.

Graphic design: Communication Division, UiB / Print: Skipnes Kommunikasjon AS

uib.no

ISBN: 9788230844915 (print)
9788230842256 (PDF)