Longitudinal Assessment of DNA Repair Signature Trajectory in Prodromal versus Established Parkinson's Disease

Danish Anwer[1], Nicola Pietro Montaldo[2,3], Hilde Loge Nilsen[2,3], Annikka Polster[1,2]

1) Division of Systems and Synthetic Biology, Department of Life Sciences, Chalmers University of Technology, 41296 Gothenburg, Sweden
2) Department of Microbiology, Oslo University Hospital and University of Oslo, 0424 Oslo, Norway
3) Cresco Centre for Embryology and Healthy Development, 0424 Oslo, Norway

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disorder characterized by motor and non-motor symptoms. DNA repair dysfunction and integrated stress response (ISR) dysregulation have been implicated in PD, but their role during the prodromal phase, before motor symptoms show, remains unclear. In this study, we analyzed longitudinal blood transcriptomic data from the Parkinson's Progression Markers Initiative (PPMI) to investigate how DNA repair and ISR gene expression differs across healthy individuals, prodromal PD patients, and individuals with established PD. Using a logistic regression classifier, we found that DNA repair and ISR gene expression effectively distinguished prodromal PD from healthy individuals, with classification accuracy increasing over time and peaking at later prodromal stages. In contrast, these pathways failed to differentiate established PD from healthy controls, suggesting that DNA repair and ISR dysregulation play a more prominent role early in disease progression. Gene expression variability was high in prodromal PD at baseline but reduced over time, indicating a convergence in gene expression patterns as the disease advances. Notably, 50% of DNA repair genes and 74% of ISR genes exhibited nonlinear expression patterns, with an initial increase followed by a decline, suggesting a transient adaptive response that weakens as PD progresses. Feature importance analysis identified key genes, including ERCC6, PRIMPOL, NEIL2, and NTHL1, as important predictors of prodromal PD. These findings suggest that DNA repair and ISR dysregulation contribute to early-stage PD pathology and may serve as biomarkers for early disease detection.

This study highlights the role of DNA repair and ISR pathways in the prodromal phase of PD, demonstrating their potential as early molecular indicators of disease onset. Future research should validate these findings in larger cohorts and explore their relevance for early diagnostics and intervention strategies.

Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder classically defined by motor symptoms (bradykinesia, tremor, rigidity, postural instability), which emerge only after significant neurodegeneration has occurred[1,2]. Years before this motor onset, patients often experience a prodromal phase marked by subtle non-motor features such as REM sleep behavior disorder (RBD), olfactory loss (hyposmia), constipation, depression, and anxiety.
This temporal progression from prodromal to overt PD suggests a window of opportunity to identify disease-driving molecular processes before extensive dopaminergic neuron (DANs) loss. Understanding the molecular changes and pathogenic mechanisms in prodromal PD is therefore critical, as it may allow for diagnosis before significant neuronal loss occurs, when neuroprotective interventions are likely most effective, as well as identify novel treatment targets.

Among the emerging mechanisms implicated in PD pathogenesis is genomic DNA damage and repair dysfunction. Growing evidence suggests that the accumulation of DNA lesions in neurons may play an active role in driving disease onset and progression [3,4]. DANs are metabolically active and exposed to high levels of endogenous reactive oxygen species (ROS) as by-products of dopamine metabolism and mitochondrial respiration[5,6]. PD is associated with impaired mitochondrial electron transport and mitophagy, leading to excessive ROS production that exacerbates DNA damage. Over time, this chronic oxidative DNA damage may overwhelm the capacity of repair pathways[7].
This results in the accumulation of both single-strand and double-strand breaks (SSBs and DSBs respectively), which are hazardous. In addition to these breaks, a plethora of base lesions can accumulate, each having distinct effects on transcription. In mitochondrial DNA (mtDNA), which has limited DNA repair capacity, the relevance of these might be more important. PD patients, in fact, exhibit a higher burden of somatic mtDNA mutations in stiratal nigral neurons compared to age-matched controls[8].

Cells counteract oxidative DNA damage primarily via the base excision repair (BER) pathway, which recognizes and excises oxidized bases[9]. The BER pathway is especially important in mitochondria, where it is the only fully functional DNA repair mechanism[10]. However, paradoxically, BER itself can become a source of stress if dysregulated. Our recent study using a *Caenorhabditis elegans* (*C. elegans)* PD model demonstrated that BER activity initiated by the DNA glycosylase NTH-1 leads to accumulation of repair intermediates (single-strand breaks) that trigger neurotoxicity during physiological aging. In this study, reducing NTH-1 levels was actually neuroprotective, preventing neurodegeneration[11]. This highlights that while DNA repair pathways are essential for cell survival, overactivation or lack of coordination in handling toxic intermediates, such as SSBs and DSBs, can cause harm by depleting resources or generating toxic DNA break ends. Consistently, genetic studies in humans have begun to link subtle variants in DNA repair genes to PD risk. For example, rare variants in the BER glycosylase NEIL2 (which, like NTH1, fixes oxidative base lesions) were found enriched in PD patients compared to controls[11]. Furthermore Sanders et al. demonstrated that genetic variants of the BER genes increase risk of PD in combination with pesticide known to affect mitochondrial

function [12]. Taken together, these findings indicate that oxidative DNA damage and an impaired or maladaptive repair response are key contributors to PD pathogenesis from its earliest stages.

A major challenge in studying DNA repair in PD is the inaccessible nature of the affected tissue, as one cannot easily monitor DNA damage or repair activity in relevant cell types in living patients. This has spurred interest in peripheral biomarkers that might reflect central neurodegenerative processes. Therefore, blood-based transcriptomics is of high interest in current PD research. However, prior transcriptomic studies largely focused on differentiating PD patients from controls after substantial cell death of up to 80% of relevant cell types had already occurred[2]. However, how these gene expression patterns change during the prodromal phase is largely unknown.

Given the evidence that DNA damage and repair defects are involved early in PD, a logical hypothesis is that prodromal PD patients likely exhibit distinct changes in DNA repair pathways before clinical manifestation. Detecting such changes in peripheral blood could provide a non-invasive marker of ongoing neurodegenerative processes. Additionally, comparing prodromal versus established PD could reveal whether there is an early, perhaps compensatory upregulation of repair mechanisms that later becomes dysfunctional as the disease advances.
By utilizing longitudinal transcriptomics from healthy and prodromal individuals as well as diagnosed PD patients, and applying machine learning, this study thus aims to explore the dynamic regulation of DNA repair gene expression in these groups.

Methods

Data

To examine longitudinal gene expression changes across PD progression, transcriptomic data from PPMI[13] were analyzed across three participant groups: healthy individuals, prodromal PD, and idiopathic PD at four time points: Baseline (BL), after 12 months (M12), after 24 months (M24) and after 36 months (M36). Whole transcriptome sequencing was performed on ribo- and globin-depleted RNA from PaxGene Tubes at HudsonAlpha using Illumina NovaSeq6000. RNA was processed via directional cDNA synthesis and NEB/Kapa library prep, aligned to GRCh38p12 using STAR. Data of the IR3 sequencing version was downloaded as TPM (generated using the Salmon method[14]) from https://ida.loni.usc.edu/.

Gene sets

We selected five gene sets representing distinct biological processes: 1) mitochondrial DNA damage repair (mtDNArep), 2) general DNA damage repair (DNArep), 3) the integrated stress response (ISR), 4) core Parkinson's disease-related genes (PD-core), and 5) Parkinson's disease-associated genes (PD-assoc). The DNArep, mtDNArep, and ISR gene sets were manually curated by HLN and MN, experts in DNA repair mechanisms. The PD-core set was compiled through manual curation from publicly available sources, including informational

material from the Michael J. Fox Foundation. The PD-assoc list was derived from a recent meta-analysis[15]. The lists are presented in supplementary table S1.

Data analyses

To evaluate the differential activity of biological processes represented by gene sets across healthy, prodromal, and established PD states, we performed classification analyses using logistic regression. We conducted this analysis for the respective pairwise comparisons (healthy vs. prodromal and prodromal vs. PD) across four time points: BL, M12, M24, M36. To ensure the reliability and robustness of our findings, we performed 1,000 iterations of the logistic regression on randomly split training and test sets for each gene set. In each iteration of the classifier, we randomly selected an equal number of subjects from each group to prevent class imbalance and minimize bias in the results. The data were stratified into training (80%) and testing (20%) sets, with the logistic regression model trained on the training set and evaluated on the test set. Classification accuracy was recorded for each iteration and averaged across all iterations to obtain a stable and reliable measure of model performance.

To assess the importance and relative contribution of individual genes, we extracted the genes' coefficients from the logistic regression model at each iteration. To enhance the robustness of these findings, we calculated the average coefficient value across the 1,000 iterations of each classifier. These average coefficients were then ranked based on their absolute values, with larger absolute values indicating higher importance of the gene in distinguishing between the groups.

To capture shifts in gene importance over time and visualize trends in disease progression, we analyzed the temporal dynamics of gene importance rankings across different time points. Code is provided at https://github.com/Polster-lab/Longitudinal-DNA-Repair-Parkinson

Results

Demographics and Clinical Characteristics

Participants included in this study were obtained from the Parkinson's Progression Marker Initiative[13]. The cohort consisted of healthy controls (at baseline n=192, 35.4% female), prodromal PD subjects (at baseline n=63, 20.6% female), and established PD patients (at baseline n=412, 34.5% female). Detailed age distributions across the four time points are shown in Figure 1.

Classification Results

Mitochondrial DNA Repair, DNA Repair, and ISR

Healthy vs established PD

Classification analyses assessed whether gene expression patterns related to mitochondrial DNA repair (mtDNArep), DNA repair (DNArep), or the integrated stress response (ISR) could

robustly distinguish healthy individuals from established PD. Classification accuracy was consistently low to moderate, ranging from approximately 0.50 to 0.64 across all time points (Figure 2A). These scores indicate that the classifier performed only slightly better than chance (random guessing would yield around 0.50 accuracy). Thus, peripheral blood gene expression in these biological pathways does not appear to provide clear, consistent markers for differentiating established PD patients from healthy individuals.

Classification analyses revealed that healthy individuals and PD patients could not be robustly distinguished using mtDNArep, DNArep, or ISR gene sets (Figure 2 A). Accuracy values remained low to moderate, fluctuating around 0.5-0.64 across all timepoints. No clear temporal trend was observed, suggesting that these pathways do not exhibit sufficiently distinct transcriptional changes in PD to enable reliable classification against healthy individuals.

Healthy vs. prodromal PD

In contrast, classification accuracy was generally high when distinguishing healthy individuals from prodromal PD patients for all three gene sets, except at the baseline (BL) visit for prodromal patients (Figure 2B). Accuracy progressively improved over time for the mtDNA repair gene set (peaking at 0.89 by visit M36) and reached its highest point at prodromal M24 for the ISR gene set (0.91). Notably, for most genes, the variance of expression was highest at BL and lowest at visit M24, suggesting that expression levels among individuals become more similar as the prodromal stage progresses. This decreasing variance likely contributes to the observed increase in classification accuracy over time and aligns with the hypothesis that molecular disruptions in these pathways occur early, becoming more consistent as prodromal individuals approach clinical PD diagnosis.

Prodromal PD vs. established PD

Similarly, classification accuracy between prodromal and established PD was consistently high across most visits, except at prodromal BL (Figure 2C). However, accuracy tended to show a slight decrease at later visits, suggesting some gene expression differences between prodromal and established PD may diminish as the disease progresses clinical

PD-Associated and Core PD Gene Sets

In addition, we analyzed core PD (PD-core) and PD-associated (PD-assoc) gene sets to compare their classification performance against DNA repair and ISR pathways. This allowed us to assess whether disease-specific gene sets capture distinct transcriptional patterns relative to repair mechanisms.

Healthy vs. PD

Notably, PD-core and PD-assoc gene sets also failed to robustly distinguish healthy individuals from PD patients in this peripheral blood dataset (Figure 2 D). Accuracy values remained low (0.49-0.65) across timepoints, confirming that transcriptional changes in these pathways do not provide sufficient discriminatory power for differentiating PD from healthy controls.

Healthy vs. Prodromal PD

The classification performance of PD-assoc and PD-core gene sets was high (0.65-0.87) when distinguishing healthy individuals from prodromal PD, except for prodromal BL. The PD-assoc gene set performed slightly better (Figure 2 E). Accuracy had a tendency to decrease slightly during the later visits.

Prodromal PD vs. Established PD

Similarly, prodromal vs. established PD classification revealed high accuracy levels (0.7-0.87) for most timepoints, similar to mtDNArep and ISR findings (Figure 2 F). Prodromal BL again failed to show robust classification accuracy, and the accuracy peaked at prodromal M24.

Global Gene Expression-Based Classification: Lack of Discrimination Between Healthy and PD

To determine whether any genes or gene combinations could robustly distinguish healthy individuals from established PD at any time point, we additionally conducted a classification analysis on the full dataset. Notably, classification accuracy remained low (0.53–0.67) across timepoints (Supplementary Figure S1), indicating that peripheral blood RNA sequencing does not provide reliable biomarkers for established PD.

Longitudinal gene expression

Longitudinal analysis of gene expression trajectories revealed that gene expression remained largely stable over time in both healthy controls and established PD patients, exhibiting minimal temporal variation. In contrast, prodromal PD showed greater variability, with many genes displaying the highest variance at BL and reaching their lowest variance at M24 (All expression profiles are available in Supplementary Figure 2). This pattern likely explains the lower classification accuracy at BL and the peak accuracy observed at M24.

Furthermore, in prodromal PD many genes exhibited dynamic, non-linear expression patterns, with 50% of DNA repair genes and 74% of ISR genes following biphasic trends (initial increases followed by decreases, or vice versa). These findings support the hypothesis that the prodromal phase is marked by an active transcriptional response, which progressively deteriorates and ultimately fails as the disease transitions into the established phase.

Feature importance

To better interpret the classification results, we extracted feature importance scores from M12, M24, and M36, ranking genes by their absolute feature importance. BL rankings were not evaluated due to low classification accuracy. Here, we report the top 10 genes in the mtDNA_rep set (27 genes total) and the top 20 genes in the DNA_rep (96 genes total) and ISR (128 genes total) sets. Table 1 further details the top 5 genes for each time point across the three sets.

mtDNA_rep

In the mtDNA_rep gene set, classifying healthy vs prodromal PD, the genes *ERCC6* and *PRIMPOL* consistently ranked in the top 10 in all three timepoints. *NEIL2, NTHL1, RAD23A* and *DNA2* ranked top 10 in M12 and M24, while *MUTYH* ranked top 10 in M24 and M36.

When classifying prodromal vs established PD, *NEIL2* and *ERCC6* consistently ranked in the top 10 in all three timepoints. *ERCC2* and *NTHL1* ranked top 10 in M12 and M24. Notably, *NTHL1* sank to rank 26 in M36. *DNA2* ranked top 10 in M24 and M36.

All rankings are shown in supplementary tables S2 and S3.

DNA_rep

In the DNA_rep gene set, classifying healthy vs prodromal PD, the genes *POLD4, XPC, POLE3* and *MAD2L2* consistently ranked in the top 20 in all three timepoints. *H2AX, ERCC6, NEIL2, APEX1, GTF2H1* and *APTX* ranked top 20 in M12 and M24, and *POLB, PRKDC, REV3L* and *POLE4* ranked top 20 in M24 and M36.

When classifying prodromal vs established PD, *POLD4, POLE4* and *ERCC6* consistently ranked in the top 20 in all three timepoints. *RIF1, MAD2L2, REV1, APEX1, POLA1, H2AFX* and *MMS19.* Notably, *APEX1, POLA1, H2AFX* and *MMS19* dropped in rank substantially in M36. *RAD23B, MBD4, PRKDC* and *REV3L* were in the top 20 in both M24 and M36. All rankings are shown in supplementary tables S4 and S5.

ISR

In the ISR gene set, classifying healthy vs prodromal PD, the genes *KDM6B, CEBPB, CDC42, S100P, HLA-DRB1, LITAF, CARS2, ATF2, IRF7, PSEN1, ERVW-1, ATF4, NFE2L2, SLC38A2, CREBBP, LARS1, CXCL8, IARS1* and *ATF6* consistently ranked in the top 20 in all three timepoints. *PTGS2, CARS1, CREB1 and RPS6KA3* ranked top 20 in M12 and M24, whereas *IARS1* ranked top 20 in M24 and M36.

When classifying prodromal vs established PD, *NFE2L2, CXCL8, CEBPB, ATF6* and *IARS1* consistently ranked in the top 20 in all three timepoints. *POLR2C, CARS1, BBC3* and *CSF1R* ranked top 20 in M12 and M24, whereas *KDM6B, STAT3, RPS6KA3, DISC1* and *GRIN2A* ranked top 20 in M24 and M36. All rankings are shown in supplementary table S6 and S7.

Discussion

This study investigated longitudinal changes in DNA repair and integrated stress response gene expression during the prodromal stages of PD using peripheral blood transcriptomics. Previous studies on DNA repair dysfunction in PD have predominantly relied on cross-sectional analyses conducted after clinical diagnosis, largely overlooking the critical prodromal period. Consequently, little is known about how gene expression patterns associated with DNA repair dysfunction and ISR evolve before clinical symptoms emerge and how these early molecular

changes relate to disease progression. Understanding these dynamic transcriptional changes in the prodromal phase is crucial, as it could reveal early biomarkers of PD and provide insights into potential compensatory mechanisms activated prior to extensive neuronal loss.

Our findings add significant nuance to the evolving understanding that disruptions in the expression of DNA repair genes occur long before PD symptoms manifest. This aligns closely with previous cross-sectional evidence of DNA damage and impaired genome maintenance mechanisms in PD patients[16]. Addressing this important gap, our longitudinal analysis provides novel insights into the evolution of DNA repair dysfunction and integrated stress response (ISR) gene expression throughout the prodromal stages of PD. Utilizing rigorous classification analyses on peripheral blood transcriptomic data, we identified that gene sets related to DNA repair and ISR effectively distinguished prodromal PD individuals from healthy controls and established PD. Conversely, these gene expression signatures did not reliably differentiate patients with clinically established PD from healthy individuals, underscoring that the most pronounced peripheral signals of DNA repair and ISR occur early in disease progression, well before classical motor symptoms manifest.

Our data further highlight substantial transcriptional variability early on in prodromal PD, and, importantly, nonlinear expression patterns in many genes. Approximately half of genes involved in DNA repair pathways and nearly three-quarters of ISR-associated genes demonstrated biphasic expression trajectories. These complex temporal patterns may reflect an initial adaptive response aimed at mitigating accumulating genomic stress, followed by an eventual decline as the compensatory mechanisms become insufficient with disease advancement.

Feature importance analysis further strengthened our findings by consistently identifying specific DNA repair genes, most importantly ERCC6, PRIMPOL, NEIL2, and NTHL1, as significant and recurring molecular predictors of prodromal PD. ERCC6 and NEIL2 consistently ranked highly across multiple assessments, emphasizing their potential role as robust early biomarkers. Interestingly, NTHL1 emerged as a strong early important feature but declined in importance at later stages, suggesting an important role in the initial prodromal phase, which may reflect an eventually overwhelmed compensatory repair mechanism as disease progresses.

ERCC6, also known as CSB, is crucial for transcription-coupled nucleotide excision repair (TC-NER), targeting actively transcribed genes for repair. It remodels chromatin to facilitate access for repair enzymes and interacts with key factors such as P53 and RNA Polymerase II[17,18]. ERCC6 dysfunction results in defective oxidative DNA repair, contributing to nerve cell death, photosensitivity, and premature aging. Mutations in ERCC6 cause Cockayne syndrome type II (CS-II), a disorder characterized by neurodevelopmental deficits and premature aging[19]. Importantly NER importance is evidenced as Ercc1-mediated DNA repair is necessary for preservation of dopaminergic neurons[20]. Mouse mutants with mild Ercc1 defects display signs of dopaminergic pathology and PD patients' peripheral cells exhibit inefficient nucleotide excision repair[20].

NEIL2, a DNA glycosylase involved in BER, plays a crucial role in maintaining genomic integrity in both nucleus and mitochondria[10]. NEIL2 associates with RNA polymerase II and other repair

proteins, suggesting its involvement in transcription-coupled repair[21]. Iron and copper, which accumulate in neurodegenerative diseases, inhibit NEIL2 activities, potentially exacerbating oxidative genome damage[22]. This inhibition can be reversed by metal chelators, including curcumin, indicating therapeutic potential[22].

PrimPol, is a recently discovered primase-polymerase, plays crucial roles in both nuclear and mitochondrial DNA maintenance[23]. It functions as a DNA polymerase, capable of extending primers and bypassing various lesions, and as a primase, catalyzing DNA primer formation and is essential for efficient DNA replication in both the nucleus and mitochondria[23,24]. It is specifically required for replication reinitiation after mtDNA damage[25]. PrimPol's ability to start DNA chains with deoxynucleotides and bypass common oxidative lesions makes it unique among primases[26]. Given its critical role in bypassing oxidative DNA lesions during mtDNA replication, PrimPol may play a key role in the cellular response to oxidative stress in prodromal PD.

Interestingly, NTHL1 initially emerged as an early prodromal marker, but its relevance sharply declined with disease progression. This is in line with a potentially critical compensatory DNA repair mechanism occurring early in disease pathogenesis and then failing. The *C. elegans* ortholog NTH-1 has been implicated in the age-dependent accumulation of single-stranded DNA breaks, contributing to PD and Alzheimer's disease (AD) pathology in *C. elegans*. In these disease models, NTH-1 generates strand breaks during physiological aging and its loss is associated with a protective phenotype characterized by reduced proteotoxicity and activation of cellular defenses that improve overall health as well as cognitive function[11,27]. Recent studies have also linked NTHL1 DNA repair to mitochondrial fitness and adaptive stress responses in human cells[28].

Collectively, our findings clearly demonstrate that dynamic changes in DNA repair dysfunction and ISR pathways occur early in PD pathogenesis and can be reliably captured in peripheral blood, highlighting their promising utility as early-stage molecular biomarkers. Prior research highlighted oxidative DNA lesions and transcription-blocking lesions as integral features of PD pathogenesis, reinforced by observations from cell models and animal studies. The longitudinal transcriptomic changes we report here extend these insights, suggesting that the DNA repair perturbations dynamically evolve as individuals transition from prodromal stages to established PD.

One compelling observation in our analysis is that, in contrast to established PD, prodromal individuals exhibit substantial gene expression volatility and non-linear trajectories, indicating an active biological response aimed at compensating for accruing molecular stress. The biphasic expression patterns observed in half of DNA repair genes during prodromal stages, which were absent in established PD, support this interpretation. We propose that this adaptive transcriptional flexibility reflects a transient attempt to counteract oxidative and genomic stress, aligning with experimental evidence of a compensatory phase before symptom onset. As PD progresses, this response may become overwhelmed, ultimately shifting to diminished homeostatic equilibrium.

Our study has several notable strengths. The longitudinal design enabled us to capture dynamic gene expression changes related to DNA repair dysfunction and ISR during prodromal PD, a perspective often missed by cross-sectional studies. Using the high-quality PPMI dataset, combined with rigorous statistical validation through extensive cross-validation, enhances the reliability and reproducibility of our findings. Additionally, the comprehensive evaluation of multiple biologically relevant gene sets ensured hypothesis-driven analyses, thereby minimizing the potential for statistical biases associated with exploratory analyses or selective reporting of findings.

However, several limitations warrant consideration. Our relatively small sample size may limit the detection of subtle transcriptional changes and reduce the accuracy in capturing disease heterogeneity. Additionally, gene expression in whole blood is moderately correlated with gene expression in the brain, suggesting peripheral blood samples may be a useful but imperfect proxy for central nervous system processes[29]. Further, variability in gene expression could also be influenced by external factors such as immune status, medication use, or comorbid conditions. Importantly, transcript levels do not directly correlate to protein levels, thus do not provide insights into all potentially relevant factors. Lastly, the demographic characteristics of our cohort potentially limit the generalizability of results.

Despite these limitations, our study contributes significantly to understanding the molecular dynamics of prodromal PD and presents several important areas that merit further investigation. Given the promising biomarkers identified, future research should validate their diagnostic potential in larger, more diverse cohorts, ideally incorporating multi-center studies to enhance generalizability, and further explore the biological mechanisms at play with *in vitro* or *in vivo* models of PD. Additionally, integrating multi-omics approaches, such as proteomics, metabolomics, or epigenomics, could deepen our understanding of DNA repair dysfunction and ISR pathways in prodromal PD and increase mechanistic understanding. Lastly, research exploring interventions targeting the early compensatory mechanisms identified through our longitudinal data could lead to novel therapeutic strategies for delaying PD progression.

In conclusion, this longitudinal study offers novel insights into the dynamics of DNA repair dysfunction and ISR gene expression in prodromal PD. Our findings highlight distinct molecular changes that occur prior to clinical diagnosis, identifying specific genes as promising biomarkers for early detection. While acknowledging its limitations, this study significantly advances our understanding of early-stage PD pathology and lays the groundwork for future research to better explore the variance and heterogeneity in prodromal PD, ultimately advancing clinical applications.

REFERENCES

1. Poewe, W. *et al.* Parkinson disease. *Nat. Rev. Dis. Primer* **3**, 1–21 (2017).

2. Cheng, H.-C., Ulane, C. M. & Burke, R. E. Clinical progression in Parkinson disease and the neurobiology of axons. *Ann. Neurol.* **67**, 715–725 (2010).

3. Welch, G. & Tsai, L.-H. Mechanisms of DNA damage-mediated neurotoxicity in neurodegenerative disease. *EMBO Rep.* **23**, e54217 (2022).
4. Martin, L. J. DNA damage and repair: relevance to mechanisms of neurodegeneration. *J. Neuropathol. Exp. Neurol.* **67**, 377–387 (2008).
5. Berman, S. B. & Hastings, T. G. Dopamine oxidation alters mitochondrial respiration and induces permeability transition in brain mitochondria: implications for Parkinson's disease. *J. Neurochem.* **73**, 1127–1137 (1999).
6. Ni, A. & Ernst, C. Evidence That Substantia Nigra Pars Compacta Dopaminergic Neurons Are Selectively Vulnerable to Oxidative Stress Because They Are Highly Metabolically Active. *Front. Cell. Neurosci.* **16**, 826193 (2022).
7. SenGupta, T. *et al.* Base excision repair causes age-dependent accumulation of single-stranded DNA breaks that contribute to Parkinson disease pathology. *Cell Rep.* **36**, (2021).
8. Coxhead, J. *et al.* Somatic mtDNA variation is an important component of Parkinson's disease. *Neurobiol. Aging* **38**, 217.e1-217.e6 (2016).
9. Bordin, D. L., Lirussi, L. & Nilsen, H. Cellular response to endogenous DNA damage: DNA base modifications in gene expression regulation. *DNA Repair* **99**, 103051 (2021).
10. Akbari, M., Nilsen, H. L. & Montaldo, N. P. Dynamic features of human mitochondrial DNA maintenance and transcription. *Front. Cell Dev. Biol.* **10**, 984245 (2022).
11. SenGupta, T. *et al.* Base excision repair causes age-dependent accumulation of single-stranded DNA breaks that contribute to Parkinson disease pathology. *Cell Rep.* **36**, 109668 (2021).
12. Sanders, L. H. *et al.* Editor's Highlight: Base Excision Repair Variants and Pesticide Exposure Increase Parkinson's Disease Risk. *Toxicol. Sci. Off. J. Soc. Toxicol.* **158**, 188–198 (2017).
13. Aleksovski, D., Miljkovic, D., Bravi, D. & Antonini, A. Disease progression in Parkinson

subtypes: the PPMI dataset. *Neurol. Sci.* **39**, 1971–1976 (2018).

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

15. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

16. Sproviero, D. *et al.* Parkinson's disease patients display a DNA damage signature in blood that is predictive of disease progression. 2024.03.25.24301713 Preprint at https://doi.org/10.1101/2024.03.25.24301713 (2024).

17. Wang, X. W. *et al.* p53 modulation of TFIIH-associated nucleotide excision repair activity. *Nat. Genet.* **10**, 188–195 (1995).

18. Tantin, D., Kansal, A. & Carey, M. Recruitment of the putative transcription-repair coupling factor CSB/ERCC6 to RNA polymerase II elongation complexes. *Mol. Cell. Biol.* **17**, 6803–6814 (1997).

19. Troelstra, C. *et al.* ERCC6, a member of a subfamily of putative helicases, is involved in Cockayne's syndrome and preferential repair of active genes. *Cell* **71**, 939–953 (1992).

20. Sepe, S. *et al.* Inefficient DNA Repair Is an Aging-Related Modifier of Parkinson's Disease. *Cell Rep.* **15**, 1866–1875 (2016).

21. Chakraborty, A. *et al.* Neil2-null Mice Accumulate Oxidized DNA Bases in the Transcriptionally Active Sequences of the Genome and Are Susceptible to Innate Inflammation. *J. Biol. Chem.* **290**, 24636–24648 (2015).

22. Hegde, M. L. *et al.* Specific Inhibition of NEIL-initiated repair of oxidized base damage in human genome by copper and iron: potential etiological linkage to neurodegenerative diseases. *J. Biol. Chem.* **285**, 28812–28825 (2010).

23. Bailey, L. J. & Doherty, A. J. Mitochondrial DNA replication: a PrimPol perspective. *Biochem. Soc. Trans.* **45**, 513–529 (2017).

24. Bailey, L. J., Bianchi, J. & Doherty, A. J. PrimPol is required for the maintenance of efficient nuclear and mitochondrial DNA replication in human cells. *Nucleic Acids Res.* **47**, 4026–4038 (2019).

25. Torregrosa-Muñumer, R. *et al.* PrimPol is required for replication reinitiation after mtDNA damage. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11398–11403 (2017).

26. García-Gómez, S. *et al.* PrimPol, an archaic primase/polymerase operating in human cells. *Mol. Cell* **52**, 541–553 (2013).

27. Tiwari, V. *et al.* Loss of DNA glycosylases improves health and cognitive function in a C. elegans model of human tauopathy. *Nucleic Acids Res.* **52**, 10965–10985 (2024).

28. Hubers, L. *et al.* Functional Impact of Nth-like DNA glycosylase on Mitochondrial Dynamics. 2024.10.10.617535 Preprint at https://doi.org/10.1101/2024.10.10.617535 (2025).

29. Sullivan, P. F., Fan, C. & Perou, C. M. Evaluating the comparability of gene expression in blood and brain. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **141B**, 261–268 (2006).

Acknowledgements

Data used in the preparation of this article was obtained on 2025-02-01 from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-data-specimens/download-data), RRID:SCR_006431. For up-to-date information on the study, visit www.ppmi-info.org.
PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research, and funding partners; including PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including 4D Pharma, Abbvie, AcureX, Allergan, Amathus Therapeutics, Aligning Science Across Parkinson's, AskBio, Avid Radiopharmaceuticals, BIAL, BioArctic, Biogen, Biohaven, BioLegend, BlueRock Therapeutics, Bristol-Myers Squibb, Calico Labs, Capsida Biotherapeutics, Celgene, Cerevel Therapeutics, Coave Therapeutics, DaCapo Brainscience, Denali, Edmond J. Safra Foundation, Eli Lilly, Gain Therapeutics, GE HealthCare, Genentech, GSK, Golub Capital, Handl Therapeutics, Insitro, Jazz Pharmaceuticals, Johnson & Johnson Innovative Medicine, Lundbeck, Merck, Meso Scale Discovery, Mission Therapeutics, Neurocrine Biosciences, Neuron23, Neuropore, Pfizer, Piramal, Prevail Therapeutics, Roche, Sanofi, Servier, Sun Pharma Advanced Research Company, Takeda, Teva, UCB, Vanqua Bio, Verily, Voyager Therapeutics, the Weston Family Foundation and Yumanity Therapeutics.

Figures

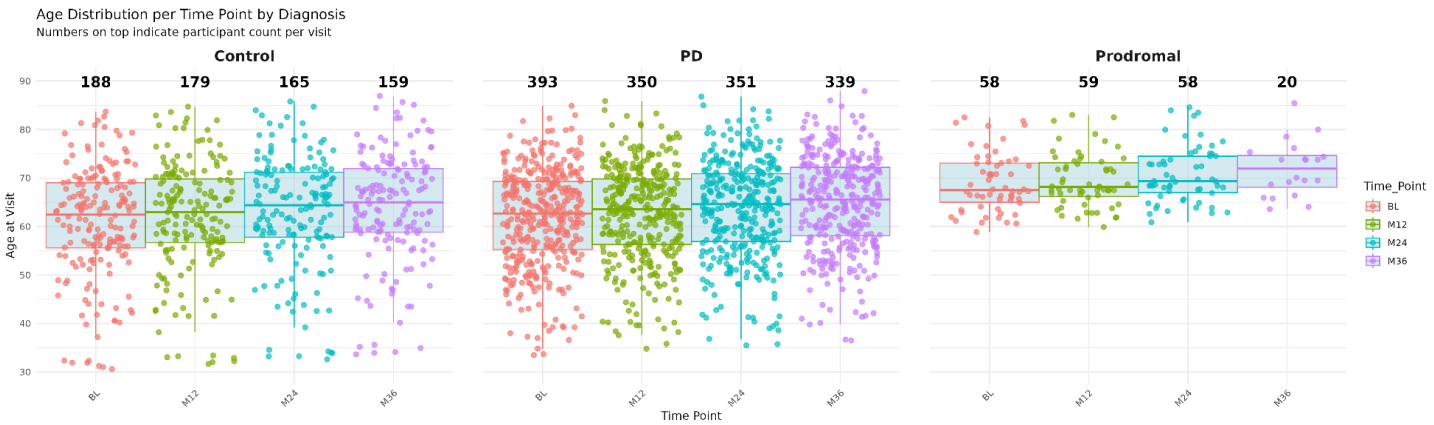

Figure 1: Age distribution across time points by diagnosis.

This figure presents the distribution of participant ages across different time points for three diagnostic groups: Healthy, established Parkinson's Disease (PD), and prodromal PD. Boxplots depict the median, interquartile range (IQR), and outliers for each group at each visit. Individual data points are overlaid, with colors indicating specific time points: baseline (BL), 12 months (M12), 24 months (M24), and 36 months (M36). The numbers above each boxplot denote the sample size at each visit.

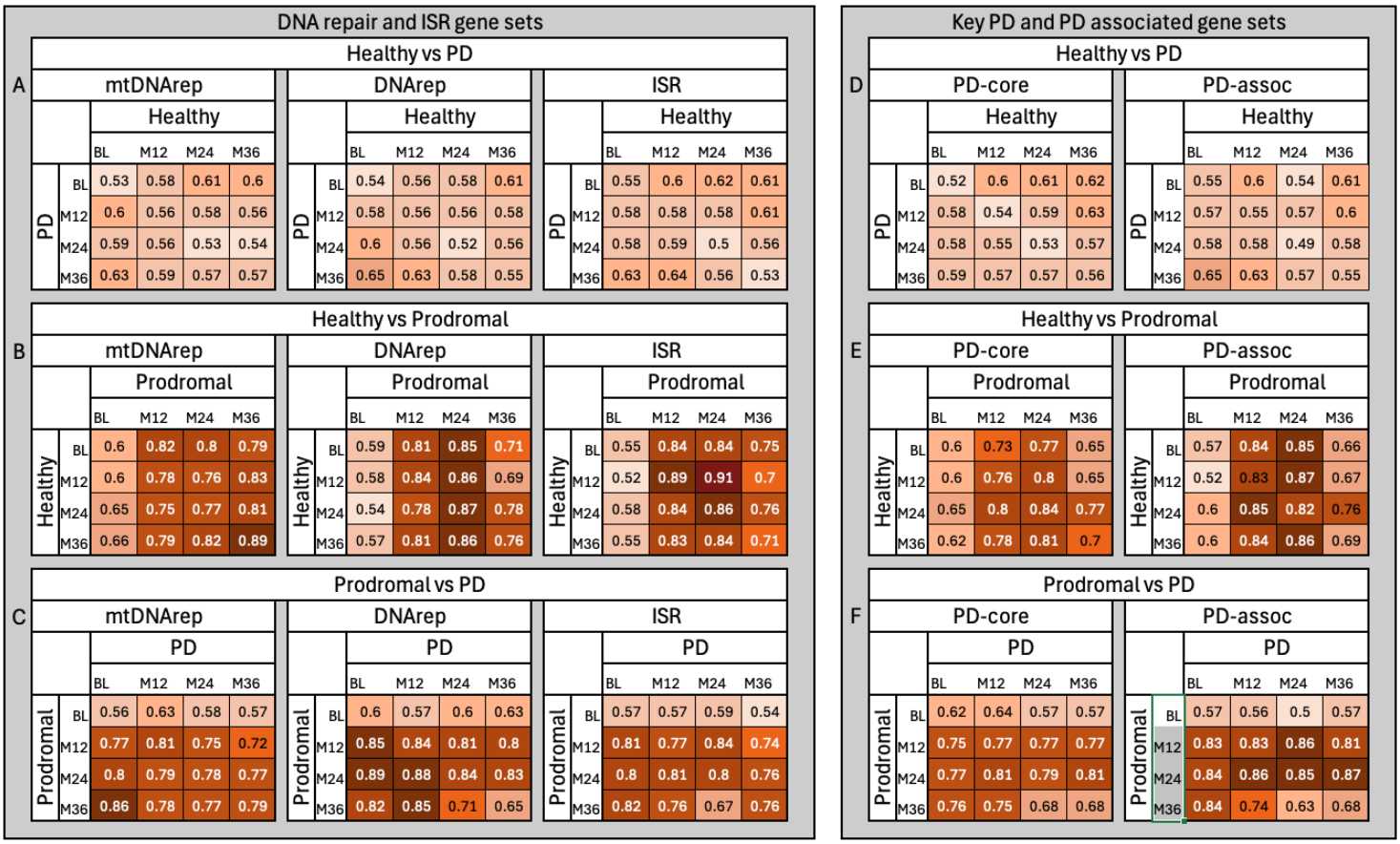

Figure 2. Classification accuracy of different gene sets in distinguishing Parkinson's disease states.

This figure presents the classification accuracy of different gene sets in distinguishing between healthy individuals ("Healthy"), those in the prodromal stage of Parkinson's disease ("Prodromal"), and individuals with established Parkinson's disease ("PD") using machine learning models. Classification accuracy, defined as the percentage of correctly classified individuals, is represented in a heatmap format, with darker shades indicating higher accuracy.

Panels (A-C) display classification results for gene sets associated with DNA repair and the integrated stress response (ISR), while Panels (D-F) show results for key Parkinson's disease-related gene sets. Accuracy is assessed at multiple time points: baseline (BL), 12 months (M12), 24 months (M24), and 36 months (M36), enabling evaluation of biomarker performance over time.

Table 1: Top five genes ranked by feature importance of the respective timepoint

mtDNA_rep						nDNA_rep						ISR					
Healthy vs Prodromal			Prodromal vs PD			Healthy vs Prodromal			Prodromal vs PD			Healthy vs Prodromal			Prodromal vs PD		
V04	V06	V08	V04	V06	V08	V04	V06	V08	V04	V06	V08	V04	V06	V08	V04	V06	V08
NEIL2	DNA2	POLB	NEIL2	DNA2	PRIMPOL	POLE3	ERCC6	POLB	POLD4	POLE3	SMUG1	CEBPB	NFE2L2	KAT2B	NFE2L2	ATF6	KDM6B
LIG3	POLQ	PRIMPOL	NTHL1	NEIL2	RAD23A	XPA	MAD2L2	RBX1	POLE3	MAD2L2	POLD4	NFE2L2	PTGS2	KDM6B	PTGS2	IARS1	KAT2B
NTHL1	NEIL2	PNKP	LIG3	NTHL1	BRCA1	H2AX	POLD4	POLD4	XPA	MMS19	POLL	CARS2	ATF6	CEBPB	POLR2C	PSEN1	NFE2L2
CRY1	NTHL1	ERCC6	MPG	ERCC6	NEIL2	POLD4	POLE	POLI	H2AFX	APEX1	POLB	CARS1	KDM6B	CDC42	NARS1	BBC3	SQSTM1
ERCC6	OGG1	PARK7	ERCC6	CRY1	DNA2	MAD2L2	PRKDC	PRKDC	APEX1	MBD4	POLE3	ATF2	CEBPB	S100P	IRF7	CEBPA	CXCL8