

<https://doi.org/10.1038/s41531-025-01238-y>

Inherent variability limits clinical utility of reproducible Parkinson's transcriptomics signatures



Roy Dayan^{1,2}✉, Serafima Dubnov³, Hagit Turm^{3,4}, Michelle Grunin^{5,6}, Shahar Shohat^{3,4}, Salim T. Khoury^{7,8}, Ami Citri^{3,4}, Tamar Harel^{2,9} & David Arkadir²

Blood transcriptomic signatures for Parkinson's disease (PD) diagnosis have failed to integrate into clinical practice despite decades of translational efforts. We evaluated the classification performance of 13 published coding RNA-based signatures using both a large public dataset and data we collected prospectively in a controlled clinical study of levodopa-naïve patients and healthy controls. Our results show that gene overlap between signatures is low but significant (2.7%, $p < 0.001$) and enriched for lipid metabolism genes. Most signatures (10/13) remained significant when tested on the Parkinson's Progression Markers Initiative (PPMI) dataset, though with lower classification performance than previously reported (median AUC: 59.7%). Performance improved for *GBA1*-associated PD. Rigorous standardization of clinical and environmental parameters in our prospective study (30 participants) failed to improve transcriptome-based classification. We conclude that while the search for a universal blood-based PD transcriptome may elucidate disease pathophysiology, its clinical utility is inherently limited.

Over the past decades, considerable efforts have been devoted to the identification of biomarkers that could enhance the diagnosis of Parkinson's disease (PD), facilitate the understanding of its pathophysiology, and assess the efficacy of potential disease-modifying therapies¹. Blood transcriptomics has emerged as a promising tool in this pursuit, offering a minimally invasive method to capture broad, unbiased molecular information². Several studies have investigated blood transcriptomics for a unique signature of PD, often using machine learning algorithms to manage high-dimensional data, and have reported success in distinguishing PD patients from healthy controls^{3–6}. Yet, despite the persistence of these efforts and the encouraging results, why have transcriptomic biomarkers not been translated into clinical practice?

Previous studies have highlighted several limitations that have hindered the translation of transcriptomics-based approaches into clinical practice. First, RNA expression exhibits substantial variability, even within homogeneous cell populations^{7,8}. Second, many transcriptomics-based classification methods and machine learning algorithms have been criticized for their susceptibility to circular reasoning⁹ and overfitting¹⁰. Consequently, gene signatures derived from blood transcriptomic data have often shown

poor concordance across studies^{4,11}. These issues raise a critical question: can PD diagnostic methods based on blood-derived gene expression signatures be reliably replicated in independent datasets, and can performance be improved by controlling parameters that contribute to variability?

To address these questions, we systematically identified published blood transcriptomics-based gene signatures that had been developed to distinguish PD patients from controls. We then evaluated the performance of these signatures by using a large, independent dataset and by conducting a prospective study designed to minimize potential confounders. Our findings show that while many signatures retain their statistically significant classification performance, the clinical applicability of these signatures remains limited—largely due to inter-individual variability, some of which can be ascribed to variance in genetic background.

Results

Poor overlap between blood transcriptomics-based signatures

A literature search, based on our predefined criteria, identified 13 studies that classified PD and healthy control subjects based on blood-derived transcriptomic signatures (Fig. 1a, Table 1). Training and validation in these

¹Department of Neurology, Hadassah Medical Center, Jerusalem, Israel. ²Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel. ³The Edmond & Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. ⁴Department of Neuroscience, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel. ⁵Braun School of Public Health, Hebrew University of Jerusalem, Jerusalem, Israel. ⁶Case Western Reserve University, Cleveland, OH, USA. ⁷Neurology unit, Nazareth Hospital EMMS, Nazareth, Israel. ⁸The Azrieli Faculty of Medicine, Bar Ilan University, Ramat Gan, Israel. ⁹Department of Genetics, Hadassah Medical Center, Jerusalem, Israel. ✉e-mail: ROYD@hadassah.org.il

13 studies were performed on 9 independent datasets (Fig. 1b). The median number of participants in each study was 205 PD patients (range: 19–400) and 189 controls (range: 19–369). 3/9 datasets included exclusively drug-naïve PD patients.

Of the 9 datasets, transcriptomic profiling was performed using microarrays in 7 and RNA-sequencing in 2 (Fig. 1b). A median of 8 genes

was identified across the 13 published signatures (range: 1–495), resulting in 411 unique genes at total after removing duplicates, non-coding transcripts, and genes for which direction of regulation was not reported. 22/411 of these genes appeared in more than one signature but only half of them (11/411, 2.7%, i.e., consensus genes) originated from at least two separate signatures that are not based on shared datasets (Table S2). While this overlap is small, it is unlikely to have occurred by chance ($p < 0.001$). None of these 11 consensus genes appeared in more than two independent signatures. Notably, all 11 consensus genes consistently showed the same direction of regulation (either up- or downregulated), indicating a degree of reproducibility.

Gene set Enrichment Analysis (GSEA) of the consensus genes revealed that 6/11 genes (*ABCA1*, *ABCG1*, *SLC25A20*, *SPON2*, *PTGDS*, and *TREM1*) were related to lipid regulation (5/11 associated with response to lipids, corrected $p = 0.012$, 3/11 associated with intracellular lipid transport, corrected $p = 0.004$, Table S3). Similarly, analysis using EnrichR revealed the strongest enrichment in lipid-related pathways.

Evaluation of classification signatures on the PPMI dataset

To assess the utility of each previously suggested signature on external datasets, we implemented a simplified scoring system in which each gene was assigned a positive or negative weight based on its originally reported direction of regulation (up- or down-regulation, respectively, see “Methods”). We first validated this approach by testing the two signatures with the largest number of genes, both derived from a single dataset (Shamir et al.; Pantaleo et al.)^{3,6}. Our scoring system successfully reproduced the ability of the original list of genes to distinguish PD from controls when applied to the relevant data sets (AUC 67% vs. reported 74% by Shamir et al. and AUC 68% vs. reported 72% by Pantaleo et al., Fig. S1). Both AUCs calculated by our scoring systems were significant ($p = 0.014$, $p < 0.001$, respectively). It is worth noting that for these two signatures, performances were similar when reported importance or effect sizes were used as weights (Shamir et al.: AUC 67.3%, $p = 0.011$; Pantaleo et al.: AUC 71.6%, $p < 0.001$) but since these metrics were only available in a subset of the studies, we opted to use equal weights to ensure consistency.

Applying our scoring system to all 13 gene signatures on the RNA-Seq-based PPMI database revealed that 10/13 achieved statistically significant

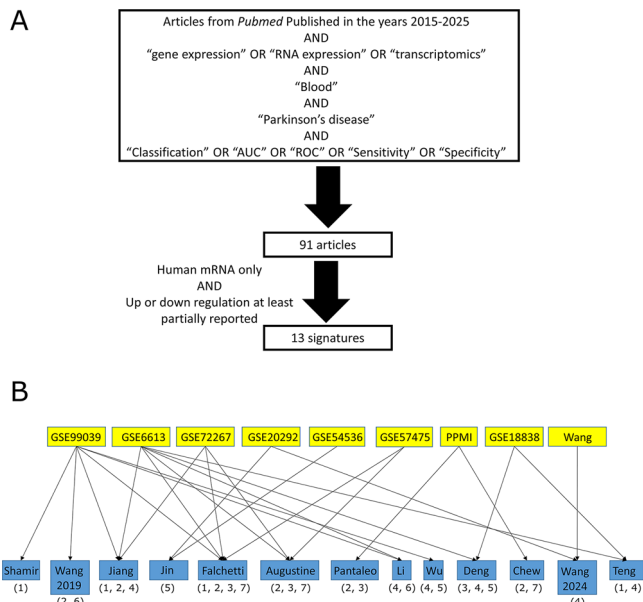


Fig. 1 | Reported signatures classifying PD and controls. **A** Flowchart of the literature search, resulting in 13 studies. **B** Original datasets (yellow boxes, upper panel) and the studies based on these datasets (blue boxes, lower panel) with the machine learning algorithms used for training the data. Digits represent the exact algorithm: (1) SVM (Support Vector Machine). (2) Random Forests. (3) Gradient Boosting. (4) LASSO (Least Absolute Shrinkage and Selection Operator). (5) Weighted Gene co-expression Network Analysis. (6) Methylation analysis. (7) Other type of analysis.

Table 1 | Classification performance and number of differentially expressed genes when applying the signatures on the PPMI and on our prospective data

Transcriptomic signature	Number of genes in the signature	Reported AUC	AUC on PPMI	AUC on prospective study $\pm$ SE	Number of genes with $ LFC > 0.1$ (same)	Number of genes with $ LFC > 0.1$ (same)
Pantaleo et al. ⁶	205 (495 ^a)	72%	68.0%†	65.6% $\pm$ 9.7%	16 (16)	3 (3)
Jiang et al. ⁵⁵	9	78%	63.3%†	58.0% $\pm$ 11.0%	1 (1)	0
Deng et al. ⁵⁶	3	74%	63.0%†	41.1% $\pm$ 10.7%	1 (1)	0
Falchetti et al. ⁴	59	84%	61.1%†	53.1% $\pm$ 10.7%	3 (3)	6 (6)
Li et al. ⁴⁷	8	76%	60.1%†	41.1% $\pm$ 10.7%	0	1 (1)
Wang et al. ⁵⁷	21	74%	59.3%†	62.1% $\pm$ 11.0%	0	3 (3)
Augustine et al. ⁵	29	82%	58.8%†	71.9%* $\pm$ 9.7%	2 (2)	5 (4)
Wu et al. ⁵⁸	7	60%	58.2%†	64.7% $\pm$ 10.5%	0	0
Shamir et al. ³	87 (100 ^a)	74%	57.1%*	48.7% $\pm$ 10.7%	1 (1)	6 (4)
Teng et al. ⁵⁹	3	80%	54.2%*	52.0% $\pm$ 10.5%	0	0
Wang et al. ⁶⁰	4	81%	52.5%	49.6% $\pm$ 11.0%	0	0
Chew et al. ⁶¹	1	56%	46.9%	50.7% $\pm$ 20.9%	0	0
Jin et al. ⁶²	1	69%	46.1%	41.3% $\pm$ 17.6%	0	0
Consensus genes	11	–	63.0%†	67.9%* $\pm$ 10.2%	1 (1)	4 (3)

*Represents $p < 0.05$; †represents $p < 0.001$.

AUC area under the curve, Reported AUC the maximum AUC reported in the original study on the test set. PPMI Parkinson's Progression Markers initiative, SE Standard Error estimated from 95% bootstrap confidence interval, LFC log-fold change, Same number of genes with the same direction of regulation in both the original study and the external dataset, PM post mortem.

^aOriginally reported including gene duplication, non-coding genes, or genes with no reported regulation direction.

classification of PD vs. controls ($p < 0.05$, Table 1). Significant signatures calculated by our scoring system had a median AUC of 59.7% (range: 54.2–68.0%). This value was lower than originally reported, with a median decrease in AUC of 15.3% (range: 1.8–25.8%, $p < 0.001$). The three signatures that did not reach significance by our scoring system contained a smaller number of genes (1–4 genes in each signature). Of note, testing the classification performance of the 11 consensus genes achieved a relatively higher AUC of 63.0% ($p < 0.001$). To assess potential confounding, we applied a logistic regression model including score, age, and sex, which yielded similar results (Table S4).

The lower classification performance calculated by our system, compared with the original studies, could be, at least partially, attributed to difference in reported gene expression between cohorts. While in the original studies, most genes were considered to be differentially expressed as defined in the PPMI (log-fold change > 0.1 or < -0.1 , e.g., 205/205 genes included from Pantaleo et al., 83/100 in Shamir et al., 59/59 in Falchetti et al.)^{3,4,6}, in the PPMI database, only 23/411 of the signature genes (5.84%) were differentially expressed. This limited overlap may be partially attributed to differences in platform technologies, as most signatures derived from expression microarray data, while PPMI employed RNA-Seq^{11,12}. Of note, all 23 differentially expressed signature genes exhibited the same direction of regulation as reported in the original reports ($p = 0.001$ for random occurrence).

Assessing the classification performance on monogenic PD from the PPMI dataset

Since the etiology of PD is heterogeneous¹³, we examined whether applying the suggested signatures on uniform genetic subgroups of PD from the PPMI database will increase the classification performance. For this analysis, we included 304 patients with negative genetic tests (idiopathic, iPD), 215 patients positive for pathogenic variants in *LRKK2* (*LRKK2*-PD), 103 patients positive for pathogenic variants in *GBA1* (*GBA*-PD), and 161 healthy controls with a negative genetic panel. We analyzed the 10/13 signatures that exhibited a statistically significant classification performance.

Limiting the classification to only *LRKK2*-PD vs. controls improved performance in 7/10 signatures, although this overall improvement was not statistically significant relative to the ability of these signatures to distinguish iPD from controls (*LRKK2*-PD vs. controls: median AUC 61.3%, range 51.1–65.2%; iPD vs. controls: median AUC 57.1%, range 54.3–68.1%; $p = 0.160$, Fig. S2). In *GBA*-PD, 9/10 signatures did better in distinguishing PD from controls (*GBA*-PD vs. controls median AUC 65.4%, range: 54.4–72.7%; iPD vs. controls: median AUC 57.1%, range: 54.3–68.1%, $p = 0.006$). Improved performance in distinguishing *GBA*-PD from controls was also noticed in the signature that was based on data from which monogenic PD were excluded⁶. With the 11 consensus genes, AUC was 71.2% in *GBA*-PD vs. controls as comparison to 63.8% in iPD vs. controls.

Although this study does not focus on asymptomatic carriers, it is noteworthy that 8/11 of the consensus genes are differentially expressed between *GBA*-PD and unaffected *GBA1* mutation carriers, and 7/11 are differentially expressed between *LRKK2*-PD and unaffected *LRKK2* mutation carriers ($p < 0.001$ for both comparisons; hypergeometric test was applied to overcome the effect secondary to different sample sizes). The direction of expression change in all of these genes (comparing monogenic PD to non-manifesting carriers) was consistent with that observed in the original studies comparing PD to healthy controls (Fig. S3).

Source of gene expression variability in the PPMI dataset

Previous studies have suggested that the classification performance of blood transcriptomics-based signatures is limited, in part, by the variability of gene expression¹¹. This variability may rise from either intra-individual differences (within the same individual) documented in repeated sampling or inter-individual differences (between different individuals), and in both cases can be affected by environmental conditions. While most environmental factors are difficult to control or quantify, the use of anti-Parkinsonian therapy represents a measurable and clinically relevant

variable. Indeed, the use of dopamine-replacement therapy has been hypothesized to influence gene expression independently of PD presence, potentially confounding transcriptomic signatures^{14,15}. Notably, 11/13 gene signatures analyzed in our study were derived from cohorts in which at least part of the participants received dopaminergic therapy.

In order to assess the potential contribution of intra- and inter-individual differences and exposure to dopaminergic therapy to variability of gene expression, we computed the mean pairwise distance (MPD) between individuals with PD and between healthy controls prior to treatment initiation. We also calculated the average gene expression change within each PD patient over the year before and after starting therapy, and for controls over equivalent time intervals.

Among all signature genes with sufficient counts (282/411), MPD values were similar between PD patients and controls ($p = 0.951$; Fig. 2a), suggesting comparable variability between individuals in both groups. As expected, MPD values within individuals across two time points were significantly lower ($p < 0.001$ for both PD and controls), reflecting lower variability within each individual across different time points, even with anti-Parkinsonian therapy. Importantly, the gene expression change following the initiation of anti-Parkinsonian therapy in PD patients did not exceed the natural variability observed across time points in untreated controls ($p = 0.960$). These findings suggest that the dominant source of variability in blood gene expression lies between individuals, rather than within individuals over time, and that long-term dopaminergic therapy does not substantially increase variability in gene expression profiles.

Intervention study attempting to reduce gene-expression variability

We hypothesized that controlling environmental factors and reducing clinical and demographic differences between individuals would reduce the variability of gene expression and improve the performance of the signatures. In addition, this may reduce some inconsistencies observed in gene regulation between cohorts previously described, also for other blood transcriptomics signatures¹¹. Towards this aim, we performed a prospective clinical study (30 participants, 16 patients with PD and 14 healthy controls, see Table S5 for demographic characteristics). We minimized factors known to affect blood transcriptome by recruiting only patients at the same clinical stage, i.e., patients naïve to symptomatic therapy required to initiate treatment¹⁶, narrowing the age range (50–80 years)^{17,18}, sampling blood at the same time of the day (7–8 AM)¹⁹ after a night fast²⁰, and requesting participants to avoid caffeinated beverages for at least 12 h²¹. The neutrophil-to-lymphocyte ratio, a potential confounder²², reported to differentiate PD from controls²³, was later calculated, and showed no significant difference between groups (PD: 1.55 ± 0.58 , Controls: 1.27 ± 0.34 , $p = 0.118$). All participant demographics are provided in Table S5.

In contrast to our hypothesis, the expression variability between individuals did not decrease neither in the PD (MPD/mean: $p = 0.813$; Coefficient of variation: $p = 1.000$) nor in the control groups (MPD/mean: $p = 0.927$; Coefficient of variation: $p = 1.000$). Similar to our results that were based on the PPMI dataset, also here only 20/411 (4.87%) were differentially expressed in PD vs. controls, 18/20 (90.0%) of them with the same direction of regulation (Table 1). As neither the variability nor the rate of differential expression improved in our cohort, classification performances were mostly similar to the results from the PPMI (in our cohort: median AUC 52.0%, range: 41.1–72.9%; PPMI: median AUC 58.8%, range: 46.1–68.0%; median difference between AUCs 3.0%, $p = 0.216$; Table 1, Fig. S4).

To examine whether levodopa induces short-term changes, in either PD or healthy individuals, 16 of the participants (6 with PD and 10 healthy controls) received a single oral dose of 125/12.5 mg levodopa/carbidopa for the first time in their life, and peripheral blood was collected for RNA sequencing 150 min later. Consistent with the long-term findings from the PPMI cohort, gene expression variability, measured by MPD, was significantly higher between individuals than within individuals (i.e., before and after levodopa intake) at baseline ($p < 0.001$ for both PD and controls; Fig. 2b). Interestingly, MPD values after levodopa were smaller in PD

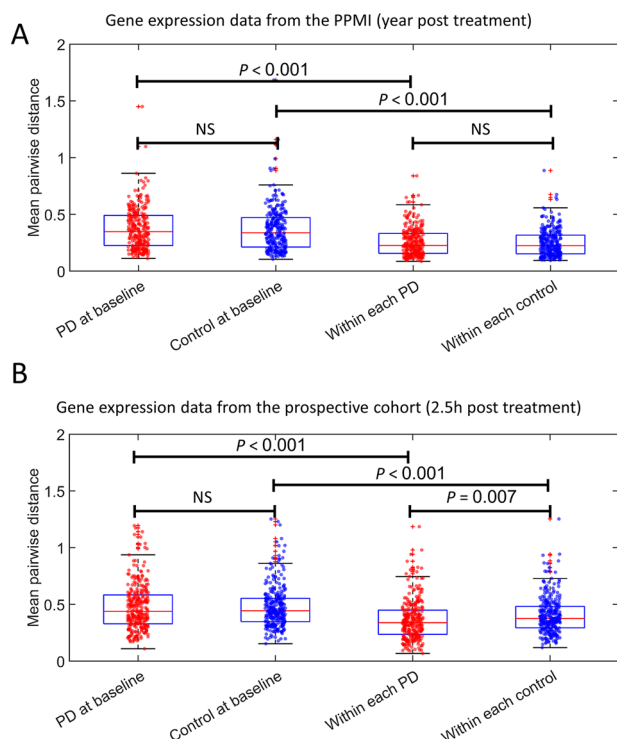


Fig. 2 | Mean pairwise distance (MPD) of gene expression values from blood RNA-seq across signature genes. **A** MPD of data from the PPMI of 282 genes, calculated between treatment-naïve PD and controls and across two time points—in the year before and in the year after initiating anti-parkinsonian therapy for PD patients, and at matched time intervals for controls. MPD was significantly lower within individuals than between individuals (PD at baseline: median 0.35 [0.11–1.45], within each PD: median 0.23 [0.09–0.84], $p < 0.001$; Controls at baseline: median 0.34 [0.10–1.68], within each control: median 0.22 [0.09–0.89], $p < 0.001$). No significant difference was observed between the treatment-related change in PD patients and the natural variation in controls ($p = 0.960$), suggesting that long-term anti-parkinsonian therapy does not substantially increase gene expression variability. **B** MPD of data from an independent prospective cohort of 16 drug-naïve individuals (6 PD, 10 controls) of 299 genes. MPD was calculated among participants before receiving treatment (baseline) and within each participant 150 min after receiving 125/12.5 mg of levodopa/carbidopa. In both PD patients and controls, variability within individuals across time points was significantly lower than variability between individuals at baseline (PD at baseline: median 0.44 [0.11–1.19], within each PD: median 0.34 [0.07–1.18], $p < 0.001$; Controls at baseline: median 0.44 [0.15–1.25], within each control: median 0.38 [0.12–1.25], $p < 0.001$). The treatment-induced change was smaller in PD than in controls ($p = 0.007$), indicating a modest but consistent within-subject response. PPMI Parkinson's Progression Markers initiative, NS non-significant, PD Parkinson's disease.

patients ($p = 0.007$), suggesting a modest differential transcriptional short-term response to dopaminergic stimulation.

Discussion

In this study, we identified a small but significant overlap (11/411) of consensus genes across previously proposed transcriptomic signatures designed to distinguish PD patients from healthy controls. While we validated the ability of 10/13 signatures to recognize PD, AUCs were overall low when applied to external datasets (median AUC 59.7%). Improvement, to a certain degree, was observed when exclusively analyzing *GBA*-PD individuals. Inter-individual gene expression variability, a major limitation in signature performance, was not reduced by prospective attempts to optimize clinical, demographic, and environmental factors of participants. We concluded that biological differences between individuals, rather than environmental or treatment effects, are the dominant source of gene

expression variation and limit the diagnostic value of blood-based expression signatures.

It is now well established that PD-related molecular changes are not confined to the brain²⁴. PD-related transcriptomic alterations have been observed in the skin²⁵, gastrointestinal system²⁶, and as was confirmed here, also in blood. These blood transcriptomic changes were shown to mirror RNA changes observed postmortem in the striatum, underscoring their potential relevance to disease biology²⁷. While the clinical utility of blood transcriptomic changes remains limited (see below), identifying reproducible molecular alterations is essential for deepening our understanding of PD pathophysiology. In this regard, it is noteworthy that 6/11 of the consensus genes are involved in lipid transport or lipid-regulatory pathways, aligning with prior reports of PD-related disruptions in lipid metabolism and its connection to alpha-synuclein pathology^{28,29}. These genes may also serve as biomarkers for suggested lipid-targeted disease-modifying therapies for PD^{30,31}.

Although the ability of blood transcriptomics-based signatures to distinguish patients with PD from controls, as was reproduced here, is overall statistically significant, their performance is not clinically useful. Among the 10 of 13 signatures that retained statistical significance in our analysis, the highest AUC was only 68.0% which is considered to be an inadequate classifier³². While our use of an equal-weighted scoring approach may have contributed to reduced performance, even in the original publications, none of the signatures exceeded an AUC of 85%, which is considered good³² but by itself does not imply clinical utility³³ and falls short of the clinical diagnosis of PD by movement disorders specialists (AUC 92.5%)³⁴.

One plausible explanation for the limited performance and low concordance among gene signatures is the genetic and pathological heterogeneity of PD¹³. Most notably, gene signatures tended to perform better in patients with *GBA*-PD, an accentuated form of the disease^{35,36}, that is uniformly associated with alpha-synuclein pathology³⁷. Conversely, when analyzing data for *LRRK2*-PD, which is associated with a more heterogeneous pathology³⁸, there was not a significant improvement in signature performances. This supports the notion that genetic stratification may improve biomarker consistency and interpretability. Interestingly, despite the well-established differences in penetrance, molecular mechanisms, and pathological substrates between *GBA*-PD and *LRRK2*-PD, we observed that similar transcriptomic signatures are shared by these two distinguished PD types. This may suggest a shared downstream common pathway³⁹.

A key challenge in developing blood transcriptomics-based classifiers is disentangling disease-specific transcriptional signatures from confounding variables. Contrary to prior assumptions¹⁴, our findings suggest that anti-parkinsonian therapy initiated within the year preceding sampling does not substantially contribute to gene expression variability. Previous studies noted several other potential confounders - such as age, time of day, fasting status, caffeine intake, and proximity to levodopa administration^{15,17–21}. In addition, RNA is an inherently unstable molecule, and variations in collection, storage, handling, and sequencing platforms can introduce substantial differences in measurement⁴⁰. However, even in our prospective cohort, where such factors were carefully minimized, variability between individuals persisted, and classification accuracy did not improve. Future studies may benefit from stricter methodological standardization, including adjustments for blood cell sub-populations, to enhance reproducibility^{41,42}.

Finally, with the advent of highly sensitive and specific diagnostic methods such as alpha-synuclein seed amplification assays⁴³, the rationale for pursuing a universal transcriptomic signature of PD becomes increasingly questionable. Instead, leveraging the unique gene expression patterns of individual patients may enable the development of divergent biomarkers that capture personalized disease trajectories. Such biomarkers would reflect the idiosyncratic nature of disease in each individual, integrating genetic predisposition (both monogenic and polygenic), disease progression pattern (e.g., body-first versus brain-first), clinical phenotype (e.g., tremor predominant vs. akinetic-rigid), and environmental exposures. Rather than obscuring this variability, transcriptomic profiling should help anticipate

transitions along the disease course or from prodromal to early-stage PD, and support personalized approaches to prognosis and therapeutic targeting.

In conclusion, we identified a modest but statistically significant overlap among published blood transcriptomics-based gene signatures for PD, alongside substantial variability between individuals that limits their clinical applicability. While future methods—including those leveraging blood-derived mRNA from single-cells or non-coding RNAs⁴⁴ may yield more robust biomarkers, current evidence suggests that pathology-driven approaches, such as alpha-synuclein seed amplification assays, offer superior clinical potential and reliability⁴⁵. Rather than pursuing a universal PD transcriptomic signature, efforts should leverage RNA variability to develop personalized biomarkers that reflect individual disease trajectories.

Methods

Data extraction from previous studies

We searched PubMed for articles (January 2015–May 2025), including “gene expression”, “RNA expression” or “transcriptomics”, plus “blood” and “Parkinson’s disease” and at least one of: “classification”, “AUC”, “ROC”, “sensitivity” or “specificity” (Fig. 1a). Of 91 articles identified, we included studies based on human mRNA reporting gene up- or downregulation.

To enhance overlap between signatures, we focused on coding genes with reported regulation. In microarray data, if genes had multiple probes (as in Shamir et al.³), we selected the probe with the highest reported importance. We calculated random co-occurrence probability using 10,000 permutation tests. Gene enrichment was evaluated using Gene Set Enrichment analysis (GSEA)⁴⁵ focusing on top 10 overlaps in KEGG Medus gene sets, GO biological process, and Human Phenotype Ontology, plus EnrichR⁴⁶ focusing on WikiPathways 2024 Human, KEGG 2021 Human, and GO Molecular Function 2025.

Analyzing performance of signatures on external datasets

To classify participants based on the gene signatures, we introduced a simplified score that quantifies the relative expression of each gene signature in a given sample. For RNA-sequencing datasets, we normalized the raw expression matrices by calculating counts per million (CPM), and for microarray datasets by quantile normalization. Normalization was followed by log₂ transformation.

Given a normalized expression matrix $M \in \mathbb{R}^{P \times N}$ measuring N signature genes across P participants, the score s_j of the j th participant is given by:

$$s_j = \sum_{i=1}^N \frac{x_{ij} - \mu_i}{\sigma_i} \cdot r_i \quad (1)$$

Here, for each signature gene, we standardized its expression across all samples in the cohort using the z-scores, and multiplied it by the regulation factor r_i , representing up- or down-regulation of gene x_i :

$$r_i = \begin{cases} -1, & \text{if } \log_2(x_i) < 0 \\ 1, & \text{if } \log_2(x_i) > 0 \end{cases} \quad (2)$$

Where $\log_2$ = log-fold change between PD and control samples, reported in the original studies.

While weights based on log-fold change have been previously used in similar settings⁴⁷, we adopted a binary approach due to large variability between datasets (see results and previous studies⁴) and limited data on gene importance or log-fold change extractable from reported studies. Statistical analysis used 10,000 label (PD/control) permutations.

We validated our equal-weight approach by testing the performance of the two largest signatures (each from a single dataset) using the original expression data it was based on. These signatures are particularly susceptible to bias from low-importance genes. For comparison, we repeated the analysis after multiplying importance scores by +1 or −1 according to the

regulation direction. To test whether the number of differentially expressed genes in the consensus set was higher than expected by chance, we used the hypergeometric distribution separately for upregulated and downregulated genes, then combined the results using Fisher’s combined probability test⁴⁸.

Transcriptomics data were obtained from the Parkinson’s Progression Markers Initiative (PPMI)⁴⁹. Data was extracted May 2025, including differential expression results and gene expression values for individual participants. We used the earliest sample from each participant. Demographic data included age on consent, sex, diagnosis, and genetic status. Overall, we collected data of 622 PD patients and 186 controls. PD and controls had a similar female-to-male ratio (PD: 42.3% females, controls: 41.4% females, $p = 0.830$). PD patients were older (PD: 64.2 ± 10.2 years, range 32.8–93.7; controls 60.9 ± 11.6 years, range 31–84.3; $p < 0.001$). Test set and source code used in Shamir et al. were downloaded from <https://github.com/Shamir-Lab/GENEPARK>³.

To compare variability both between individuals and within individuals across two time points, we used MPD, reflecting global similarity unaffected by sample number. MPD was calculated on normalized gene expression data by averaging absolute values of all relevant combinations within groups or differences before/after treatment for each individual. MPD was assessed for all signature genes, excluding those filtered for low counts in any of the groups. Group comparisons used the Kruskal–Wallis test. General variability between individuals in different datasets was measured by MPD/mean and coefficient of variability (CV, standard deviation/mean). One-tailed Wilcoxon signed-rank test assessed whether variability decreased after controlling for environmental conditions.

Prospective study to reduce expression variability

Participants were recruited at the Hadassah Medical Center movement disorder clinic (June 2020–June 2022). All participants provided informed consent (IRB 0144-20-HMO). We included patients aged 50–80 diagnosed with PD following the MDS (Movement Disorders Society) criteria⁵⁰, naïve to symptomatic therapy (except MAO-B inhibitors), at clinical stage requiring symptomatic therapy initiation. Exclusion criteria were atypical PD, negative 6-[18 F]fluoro-L-DOPA scan, or prior neuroleptic use. Healthy controls were included after ruling out Parkinsonian signs. Baseline comparisons used Student’s t -test for age and Fisher’s exact test for sex.

PD patients and controls were tested following overnight fasting from caffeine-containing beverages and food. Upon arrival, participants were examined, the motor part of the Unified Parkinson’s disease rating scale (UPDRS III) was calculated, and 15 ml of peripheral blood for RNA-Seq was obtained. Then, some participants were treated with a tablet containing 125 mg of levodopa and 12.5 mg of carbidopa. Additional physical examination was performed, and blood samples were obtained 150 min post-dose.

Blood samples were immediately preserved in RNA stabilization tubes (Tempus™) and stored in -80°C . Total RNA was extracted using the Tempus™ Spin RNA Isolation Kit, per manufacturer’s instructions. RNA concentration and purity were assessed using NanoDrop™ 2000 spectrophotometer (Thermo Fisher Scientific), and RNA integrity was evaluated using the Agilent 2100 Bioanalyzer (Agilent Technologies).

For each sample, 100 ng of total RNA was used to prepare RNA-seq libraries using the QuantSeq 3′ mRNA-Seq Library Prep Kit, per manufacturer’s protocol. To reduce globin mRNA-derived fragments, the RS-Globin Block module was incorporated after first-strand cDNA synthesis. To eliminate amplification bias, the standard Second Strand Synthesis Mix (SS1) was replaced with the UMI Second Strand Synthesis Mix (USS). Libraries from 41 to 44 blood samples were pooled and sequenced on the Illumina NextSeq platform. Reads were aligned to the hg38 reference genome using STAR, and UMI-tools were used for UMI-based deduplication and generation of gene-level count matrices. Full reagents are in Table S1.

Data analysis used R (version 4.4.2; R Core Team, 2024) in the RStudio environment and MATLAB R2024b (The MathWorks, Inc., Natick, MA). RNA-seq counts were processed using edgeR⁵¹. Genes with expression levels below 1 CPM in fewer than three samples were filtered out, retaining 17,223

genes. Normalization used the trimmed mean of M-values (TMM) method followed by conversion to log2CPM. Correction for batch effect was performed using ComBat-seq⁵². Differential expression to compare PD and control, and time after levodopa intake was calculated using the limma R package⁵³. *P*-values were corrected via the Benjamini-Hochberg method. Cell-type deconvolution was performed using robust linear regression with non-negative constraints, using the LM22 signature matrix⁵⁴. The reference was scaled to match the distribution of TMM-normalized bulk data. Estimated proportions were normalized per sample, and neutrophil levels were compared across groups relative to calculated levels of B cells, T cells, and natural killer cells.

Data availability

Data used in the preparation of this article were obtained on [2025-05-01] from the Parkinson's Progression Markers Initiative (PPMI) database (<https://www.ppmi-info.org/access-data-specimens/download-data>), RRID:SCR_006431. For up-to-date information on the study, visit <http://www.ppmi-info.org>. Data from the prospective cohort is available upon request.

Received: 30 September 2025; Accepted: 8 December 2025;

Published online: 19 December 2025

References

- Parnetti, L. et al. CSF and blood biomarkers for Parkinson's disease. *Lancet Neurol.* **18**, 573–586 (2019).
- Chen-Plotkin, A. S. Unbiased approaches to biomarker discovery in neurodegenerative diseases. *Neuron* **84**, 594–607 (2014).
- Shamir, R. et al. Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **89**, 1676–1683 (2017).
- Falchetti, M., Prediger, R. D. & Zantotto-Filho, A. Classification algorithms applied to blood-based transcriptome meta-analysis to predict idiopathic Parkinson's disease. *Comput Biol. Med.* **124**, 103925 (2020).
- Augustine, J. & Jereesh, A. S. Blood-based gene-expression biomarkers identification for the non-invasive diagnosis of Parkinson's disease using two-layer hybrid feature selection. *Gene* **823**, 146366 (2022).
- Pantaleo, E. et al. A machine learning approach to Parkinson's disease blood transcriptomics. *Genes* **13**, 727 (2022).
- Peixoto, L. et al. How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Res.* **43**, 7664–7674 (2015).
- Eling, N., Morgan, M. D. & Marioni, J. C. Challenges in measuring and understanding biological noise. *Nat. Rev. Genet.* **20**, 536–548 (2019).
- Barbash, S. & Soreq, H. Statistically invalid classification of high throughput gene expression data. *Sci. Rep.* **3**, 1102 (2013).
- Cheng, Y., Xu, S. M., Santucci, K., Lindner, G. & Janitz, M. Machine learning and related approaches in transcriptomics. *Biochem. Biophys. Res. Commun.* **724**, 150225 (2024).
- Chikina, M. D. et al. Low-variance RNAs identify Parkinson's disease molecular signature in blood. *Mov. Disord. J. Mov. Disord. Soc.* **30**, 813–821 (2015).
- Chikina, M. D. & Sealfon, S. C. Increasing consistency of disease biomarker prediction across datasets. *PLoS ONE* **9**, e91272 (2014).
- Thenganatt, M. A. & Jankovic, J. Parkinson Disease Subtypes. *JAMA Neurol.* **71**, 499 (2014).
- Chen-Plotkin, A. S. Blood transcriptomics for Parkinson disease?. *Nat. Rev. Neurol.* **14**, 5–6 (2018).
- Smith, L. M. et al. Striatal mRNA expression patterns underlying peak dose L-DOPA-induced dyskinesia in the 6-OHDA hemiparkinsonian rat. *Neuroscience* **324**, 238–251 (2016).
- Schmitt, I. et al. L-dopa increases α -synuclein DNA methylation in Parkinson's disease patients in vivo and in vitro. *Mov. Disord.* **30**, 1794–1801 (2015).
- Wang, Z. et al. Multi-omics analysis reveals the genetic aging landscape of Parkinson's disease. *Sci. Rep.* **14**, 31167 (2024).
- Peters, M. J. et al. The transcriptional landscape of age in human peripheral blood. *Nat. Commun.* **6**, 8570 (2015).
- Zhang, J. et al. Identification of key genes and diagnostic model associated with circadian rhythms and Parkinson's disease by bioinformatics analysis. *Front. Aging Neurosci.* **16**, 1458476 (2024).
- Leonardson, A. S. et al. The effect of food intake on gene expression in human peripheral blood. *Hum. Mol. Genet.* **19**, 159–169 (2010).
- Chen, C. C. et al. Investigating plasma metabolomics and gut microbiota changes associated with Parkinson disease: a focus on caffeine metabolism. *Neurology* **104**, e213592 (2025).
- Xu K. et al. Decreased SNCA expression in whole-blood RNA analysis of parkinson's disease adjusting for lymphocytes. Preprint posted online November 19 <https://doi.org/10.1101/2024.11.18.623684> (2024)
- Muñoz-Delgado, L. et al. Peripheral immune profile and neutrophil-to-lymphocyte ratio in Parkinson's disease. *Mov. Disord.* **36**, 2426–2430 (2021).
- Liu T., Wu H. & Wei J. Beyond the brain: exploring the multi-organ axes in Parkinson's disease pathogenesis. *J. Adv. Res.* <https://doi.org/10.1016/j.jare.2025.05.034> (2025).
- Planken, A. et al. Looking beyond the brain to improve the pathogenic understanding of Parkinson's disease: implications of whole transcriptome profiling of Patients' skin. *BMC Neurol.* **17**, 6 (2017).
- Kurz, A. et al. Differential expression of gut miRNAs in idiopathic Parkinson's disease. *Park. Relat. Disord.* **88**, 46–50 (2021).
- Irmady, K. et al. Blood transcriptomic signatures associated with molecular changes in the brain and clinical outcomes in Parkinson's disease. *Nat. Commun.* **14**, 3956 (2023).
- Galper, J. et al. Lipid pathway dysfunction is prevalent in patients with Parkinson's disease. *Brain* **145**, 3472–3487 (2022).
- Kamano, S., Ozawa, D., Ikenaka, K. & Nagai, Y. Role of lipids in the pathogenesis of Parkinson's disease. *Int J. Mol. Sci.* **25**, 8935 (2024).
- Bar-On, P. et al. Statins reduce neuronal α -synuclein aggregation in in vitro models of Parkinson's disease. *J. Neurochem.* **105**, 1656–1667 (2008).
- Vincent, B. M. et al. Inhibiting stearyl-CoA desaturase ameliorates α -synuclein cytotoxicity. *Cell Rep.* **25**, 2742–2754.e31 (2018).
- White, N., Parsons, R., Collins, G. & Barnett, A. Evidence of questionable research practices in clinical prediction models. *BMC Med.* **21**, 339 (2023).
- Janssens, A. C. J. W. & Martens, F. K. Reflection on modern methods: revisiting the area under the ROC Curve. *Int. J. Epidemiol.* **49**, 1397–1403 (2020).
- Di Biase, L., Pecoraro, P. M. & Di Lazzaro, V. Validating the accuracy of Parkinson's disease clinical diagnosis: a UK brain bank case-control study. *Ann. Neurol.* **97**, 1110–1121 (2025).
- Dayan, R. et al. Atrophy-related corticospinal changes in advanced Parkinson's disease are associated with the genetic etiology of the disease. *J. Park Dis.* **14**, 1584–1593 (2024).
- Skrahn, A. et al. GBA1-associated Parkinson's disease is a distinct entity. *Int. J. Mol. Sci.* **25**, 7102 (2024).
- Clark L. N. et al. Association of glucocerebrosidase mutations with dementia with Lewy bodies. *Arch. Neurol.* **66**, 578–83 (2009).
- Kalia, L. V. et al. Clinical correlations with lewy body pathology in LRRK2-related Parkinson disease. *JAMA Neurol.* **72**, 100 (2015).
- Ysselstein, D. et al. LRRK2 kinase activity regulates lysosomal glucocerebrosidase in neurons derived from Parkinson's disease patients. *Nat. Commun.* **10**, 5570 (2019).
- Kornienko, I. V., Aramova, O. Y. U., Tishchenko, A. A., Rudoy, D. V. & Chikindas, M. L. RNA stability: a review of the role of structural features and environmental conditions. *Molecules* **29**, 5978 (2024).
- Vanhaverbeke, M. et al. Peripheral blood RNA biomarkers for cardiovascular disease from bench to bedside: a position paper from

- the EU-CardioRNA COST action CA17129. *Cardiovasc. Res.* **118**, 3183–3197 (2022).
42. Nido, G. S. et al. Single-nucleus transcriptomics reveals disease- and pathology-specific signatures in α -synucleinopathies. *Brain* **148**, 1588–1603 (2025).
 43. Okuzumi, A. et al. Propagative α -synuclein seeds as serum biomarkers for synucleinopathies. *Nat. Med.* **29**, 1448–1455 (2023).
 44. Madrer, N. et al. Pre-symptomatic Parkinson's disease blood test quantifying repetitive sequence motifs in transfer RNA fragments. *Nat. Aging* **5**, 868–882 (2025).
 45. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
 46. Chen, E. Y. et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinforma.* **14**, 128 (2013).
 47. Li, D. et al. Integrative analysis of DNA methylation and gene expression data for the diagnosis and underlying mechanism of Parkinson's disease. *Front. Aging Neurosci.* **14**, 971528 (2022).
 48. Yoon, S., Baik, B., Park, T. & Nam, D. Powerful p-value combination methods to detect incomplete association. *Sci. Rep.* **11**, 6980 (2021).
 49. Craig, D. W. et al. RNA sequencing of whole blood reveals early alterations in immune cells and gene expression in Parkinson's disease. *Nat. Aging* **1**, 734–747 (2021).
 50. Postuma, R. B. et al. MDS clinical diagnostic criteria for Parkinson's disease: MDS-PD clinical diagnostic criteria. *Mov. Disord.* **30**, 1591–1601 (2015).
 51. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
 52. Zhang, Y., Parmigiani, G. & Johnson, W. E. ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinforma.* **2**, lqaa078 (2020).
 53. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).
 54. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. in *Methods in Molecular Biology* (ed. John M. Walker) 243–259 (Springer, 2018).
 55. Jiang, F., Wu, Q., Sun, S., Bi, G. & Guo, L. Identification of potential diagnostic biomarkers for Parkinson's disease. *FEBS Open Bio* **9**, 1460–1468 (2019).
 56. Deng, H. W. et al. Revealing novel genes related to Parkinson's disease pathogenesis and establishing an associated model. *Neuroscience* **544**, 64–74 (2024).
 57. Wang, C., Chen, L., Yang, Y., Zhang, M. & Wong, G. Identification of potential blood biomarkers for Parkinson's disease by gene expression and DNA methylation data integration analysis. *Clin. Epigenetics* **11**, 24 (2019).
 58. Wu, Z. et al. A computational approach based on weighted gene co-expression network analysis for biomarkers analysis of Parkinson's disease and construction of diagnostic model. *Front. Comput. Neurosci.* **16**, 1095676 (2022).
 59. Teng, W. B. et al. Exploring and validating key genetic biomarkers for diagnosis of Parkinson's disease. *Brain Res. Bull.* **220**, 111165 (2025).
 60. Wang, L. et al. A diagnostic model for Parkinson's disease based on circadian rhythm-related genes. *J. Transl. Med.* **22**, 635 (2024).
 61. Chew, G. et al. Transcriptomic imputation of genetic risk variants uncovers novel whole-blood biomarkers of Parkinson's disease. *NPJ Park Dis.* **10**, 99 (2024).
 62. Jin, X. et al. Weighted gene co-expression network analysis reveals specific modules and biomarkers in Parkinson's disease. *Neurosci. Lett.* **728**, 134950 (2020).

Acknowledgements

RD received support from the Israeli Center for Addictions (ICA). RD and DA received support from Berel and Agnes Ginges family foundation. Initial support for the project was awarded from the Israel Ministry of Health to DA and AC (Grant 3-17303). We thank Professor Alberto J. Espay for his advice and insight. 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.

Author contributions

R.D., S.D., A.C., and D.A. designed the study. R.D., D.A., and S.T.K. were responsible for patient recruitment and clinical evaluation. H.T., S.S., and A.C. carried out the RNA extraction. R.D., S.D., M.G., A.C., T.H., and D.A. contributed to data analysis and interpretation. R.D. and D.A. drafted the manuscript and prepared the figures. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41531-025-01238-y>.

Correspondence and requests for materials should be addressed to Roy Dayan.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025