Journal of Parkinson's Disease 3 (2013) 19–29
DOI 10.3233/JPD-120159
IOS Press

Research Report

Found in transcription: Accurate Parkinson's disease classification in peripheral blood

Magdalena Kauczynska Karlsson[a,b,*], Praveen Sharma[a], Jan Aasly[c], Mathias Toft[d], Örjan Skogar[e], Solve Sæbø[b] and Anders Lönneborg[a]
[a]*DiaGenic ASA, Oslo, Norway*
[b]*Department of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Norway*
[c]*Department of Neurology, St Olavs Hospital, Trondheim University Hospital, Norway*
[d]*Department of Neurology, Oslo University Hospital, Norway*
[e]*FUTURUM, County Hospital Ryhov, Jönköping, Karolinska Institutet, Stockholm, Sweden*

Abstract.
Background: A blood-based test for the early detection of Parkinson's disease (PD) would be an important diagnostic tool and useful for patient selection when developing novel drugs or treatments for the disease.
Objective: Here, we aimed to identify potential biomarkers associated with PD.
Methods: We applied gene expression profiling to the study of peripheral blood from 75 healthy control subjects and 79 PD patients at different stages of the disease. Healthy control subjects were matched for age and gender with PD subjects, and the diagnosis of patients was based on clinical evaluation by specialists in movement disorders. RNA was extracted from the blood samples and the gene expressions were measured using the Illumina HumanHT-12 v4.0 Expression BeadChip.
Results: Our results support previous studies that gene expression in blood may be instrumental in the search for molecular biomarkers for PD. Single cross-validation results show that PD can be correctly classified from healthy controls with an agreement of 88% to clinical diagnosis. *De novo* PD patients are classified with a sensitivity of 87%, which is close to what was achieved for the patients having a confirmed PD diagnosis with disease duration <5 and >5 years (93% and 88%). A double cross-validation procedure showed that using a selected set of around 650 informative genes, similar results are achieved. Functional analysis of the selected genes showed genes significantly associated to mitochondrial dysfunction, protein ubiquitination, gene expression and cell death.
Conclusions: PD affects gene expression in blood, suggesting the potential for the development of a blood-based gene expression test.

Keywords: Parkinson's disease, biomarker, blood, classification, gene expression

INTRODUCTION

The incidence of Parkinson's disease (PD) increases with age and is estimated to be 0.1–0.2% over the age of 65 years [1] and reaching a prevalence of almost 4% in those aged above 85 years. By 2030, prevalence is expected to at least double due to an ageing population [2]. Currently, all available treatments for PD, both pharmacological and surgical, offer only symptomatic benefit. These treatments improve the quality of life for persons with PD, but they do not stop progression of the disease. A blood-based test for the early detection of PD would be important for clinical practice, but also a useful tool for patient selection when developing novel drugs or other treatments directed towards protection or disease modifying properties.

Conventional diagnosis of PD is purely clinical and based on medical history and a detailed neurological

*Correspondence to: Magdalena Kauczynska Karlsson, DiaGenic ASA, Oslo, Norway. Tel.: +47 2324 8965; Fax: +47 2324 8959; E-mail: magdalena.karlsson@diagenic.com.

ISSN 1877-7171/13/$27.50

examination using clinical scales to test for the presence of characteristic motor symptoms: tremor, bradykinesia, rigidity and postural instability due to loss of postural reflexes. These classic approaches are subjective, depend on the experience of the treating physician, and only applicable in the clinical stage of the disease. The characteristic motor symptoms are related to dopamine deficiency, a consequence of progressive loss of dopaminergic neurons in the substantia nigra and other brain structures. Because the manifestation of symptoms may take years to develop, more than 70–80% of the dopaminergic neurons will have degenerated before the diagnosis is made. Familial etiology of PD exists in approximately 5% of the affected population [3, 4] and since the late 1990s, when the first monogenic form of PD was discovered, a number of genes and loci have been implicated in the disease. These include α-synuclein (*SNCA*/PARK1/4), *parkin* (*PRKN*/PARK2), ubiquitin carboxyl-terminal esterase L1 (*UCHL-1*/PARK5), Pten-induced kinase 1 (*PINK1*/PARK6), Oncogene *DJ-1* (DJ-1/PARK7) and leucine-rich repeat kinase 2 (*LRRK2*/PARK8), with vacuolar protein-sorting associated protein 35 (*VPS35*/PARK17) [5] and eukaryotic translation initiation factor 4-gamma 1 (*EIF4G1*/PARK18) [6] being the latest genes added to the list in 2011. Analysis of the expression of genes, the proteins these genes produce and the pathways that they are involved in, may provide essential clues to our understanding of the molecular pathogenesis of PD, including dopaminergic neuronal death.

A number of molecular studies on gene expression profiles for PD in peripheral blood [7–13], have, together, shown that a significant change in gene expression can be detected for PD subjects compared to healthy control subjects and disease controls such as Alzheimer's disease, suggesting that there exists a potential to develop a blood test for prediction of early (preclinical) PD. To date, none of the described results have been validated in an independent cohort, and no blood test for PD detection based on gene expression has been developed. However, the development of a blood-based gene expression test for Alzheimer's disease has recently been published [14–16] suggesting that accurate blood-based tests can be developed for major neurological diseases.

In the current study, we performed a large-scale gene expression study using Illumina microarrays on peripheral blood samples from patients with PD to develop a prediction model based on canonical partial least squares (CPLS) and to investigate its diagnostic accuracy in detection of *de novo* PD. We additionally looked at how disease activity and treatment affected the blood gene expression profile.

MATERIALS AND METHODS

Ethics Statement

All subjects included in the study gave written informed consent, and the local ethics committees approved all procedures, including blood sample collection (Ref. No. 4.2008.1123, REC Central, Norway; Ref. No. 154-08109c 2008-4196, REC South East, Norway; DNr. M217-08, EPN Linkoping, Sweden).

Study subjects

Patients with clinically defined PD, patients with *de novo* PD and healthy control subjects matched for age and gender were recruited from three clinical centers located in Norway and Sweden from August 2008 until October 2010. Diagnostic evaluation included the UK Parkinson's Disease Society Brain Bank Criteria (UKPDSBB) [17], Hoehn and Yahr staging [18], and Unified Parkinson's Disease Rating Scale (UPDRS) [19, 20] for PD and *de novo* PD patients. Clinical diagnosis was established by an experienced movement disorder specialist who performed a clinical interview, medical examination and studied medical records, developmental history and other available diagnostic information. Control subjects were healthy individuals with no apparent neurological symptoms, recruited as age- and gender-matched controls to PD and *de novo* PD subjects. The majority of the healthy control subjects were spouses. Exclusion followed if other neurological disease, i.e. epilepsy, multiple sclerosis and dementia, or severe depression was present. Dementia was evaluated based on a cognitive evaluation consisting of Mini–Mental State Examination (MMSE) score [21], clinical interview and the use of UPDRS, which includes a question about cognition. Depression was evaluated based on the clinical interview and UPDRS, which contains a question regarding depression. No additional tests were conducted to diagnose depression. Demographic and clinical information was obtained for all patients and control subjects included in the study (Tables 1, 2).

The inclusion of patients with a known PD diagnosis was based on: (1) a diagnosis of PD according to the UKPDSBB criteria; (2) a PD diagnosis for more than five years after onset; (3) on medical treatment for PD; and (4) MMSE equal to or greater than 27. Inclusion of patients with newly diagnosed PD was based on: (1) a

Table 1
Demographic characteristics of Parkinson's disease patients and healthy control subjects

	PD subjects $n = 79$ Mean (Range)	Healthy control subjects $n = 75$ Mean (Range)	All subjects $n = 154$ Mean (Range)
Demographics			
Age (years)	65.0 (43–83)	63.3 (35–80)	64.2 (35–83)
Gender (male)	51.9%	49.3%	50.6%
Education (years)	13 (7–23)	13 (7–21)	13 (7–23)

Table 2
Characteristics of Parkinson's disease patients.

Score		Mean	Median	SD	Min	Max
De novo PD	Hoehn and Yahr	1.67	2	0.5	1	2.5
	UPDRS	31.2	31	9.5	15	49
Treated PD*	Hoehn and Yahr	2	2	0.6	0	3
	UPDRS	31.4	32	12.7	11	62
All PD	Hoehn and Yahr	1.9	2	0.6	0	3
	UPDRS	31.3	32	11.8	11	62

*_De novo_ PD patients were examined before they started taking dopaminergic drugs. Treated PD patients were examined in the on-medication state. This is probably the reason for the similar UPDRS scores in the two groups.

PD diagnosis for less than five years after onset; (2) on medical treatment for PD; (3) Hoehn and Yahr equal to or less than three; and (4) MMSE score equal to or greater than 27. *De novo* PD subjects were recruited based on: (1) no final PD diagnosis (2) no medical treatment for PD; (3) Hoehn and Yahr equal to or less than three; and (4) a clinical picture assuming development of PD. The *de novo* PD subjects subsequently received dopaminergic therapy and their condition was followed up annually for up to 3 years past first visit to confirm diagnosis. Four of the *de novo* PD patients were followed up once within 3 years past first visit. Healthy control subjects were mostly recruited from spouses or family of patients to the study based on: (1) MMSE score greater than 27; (2) age-matched as cohort to patients; and (3) no signs or symptoms suggesting neurological disease.

Blood sampling

Blood sampling was performed at the same time as the clinical evaluation. Venous blood samples (2.5 mL) were drawn into PAXgene™ tubes (Becton & Dickenson, Qiagen Inc., Valencia, CA) according to the manufacturer's instructions. Tubes were incubated at room temperature (18–25°C) overnight prior to freezing and storage at −70°C or below. Tubes were transported on dry ice to DiaGenic's laboratory in Oslo, Norway, and stored at −70°C or below until processed further. RNA was extracted from all samples within 6 months of blood draw.

RNA extraction and Quality Control

The blood samples were thawed for 2 hours before total RNA was extracted using PAXgene™ Blood RNA kit (Qiagen Inc., Valencia, CA) according to the manufacturer's instructions. Total RNA was stored at −70°C or below until analysis. The RNA was assessed for quality and quantity using the NanoDrop ND-1000 spectrophotometer (NanoDrop, Wilmington, DE) and the Agilent 2100 BioAnalyzer (Agilent, Santa Clara, CA), with sample acceptance limits RIN $\geq$ 7.3; 28 S/18 S $\geq$ 1.0; A260/A230 $\geq$ 1.0; A260/A280 $\geq$ 1.8; and RNA concentration $\geq$15 ng/L. cDNA was prepared in batches using the High-Capacity cDNA Reverse Transcriptase kit (Applied Biosystems, Foster City, CA) according to manufacturer's instructions.

Microarray experimental design

All samples were organized in batches of 12 due to microarray experimental steps. Each batch consisted of five PD subjects and five healthy control subjects, together with two technical replicates used solely for quality control assessment. All patients and control subjects were distributed so that each batch contained five women and five men, equally distributed between disease status, and a total of 17 batches were randomly created as described.

RNA samples were shipped on dry ice to the AROS Applied Biotechnology laboratories (Aarhus, Denmark) for microarray analysis, and the study was blinded for the operators.

Microarray procedure

The gene expression screen was performed using the Illumina whole-genome expression array HumanHT-12 v4.0 Expression BeadChip (Illumina, Inc., San Diego, CA), containing 47, 231 oligonucleotide probes representing 34,602 genes, and processed according to the manufacturer's protocol. The Illumina® TotalPrep™ RNA Amplification Kit (Applied Biosystems, CA) was used to prepare labeled cRNA. To

start, 600 ng of total RNA were used. For the synthesis of cDNA, T7-oligo (dT) primers were used and the cDNA then underwent second strand synthesis before it was purified and subjected to an *in vitro* transcription (IVT) labeling using T7 RNA Polymerase. The labeled and fragmented cRNA was hybridized overnight on the BeadChip (Illumina, Inc., San Diego, CA). The arrays were then washed, blocked, and stained using streptavadin-Cy3. The arrays were finally scanned on an Illumina BeadArray Reader following the manufacturer's protocols. Illumina BeadStudio software (Illumina, Inc., San Diego, CA) was used to quantitatively detect fluorescence emission by Cy3, and generate signal intensity values, detection *p*-values, average bead number and bead standard deviation from the scans.

Microarray data quality control

To ensure high sample signal quality the following quality criteria were applied on the microarray data: (1) homogenous background signal based on visual inspection; (2) the average of the average signal on each microarray to be within 100–150, and the average signal on all microarrays to be within 20% of the overall average; (3) the average signal/noise ratio to be above ten on each microarray; (4) the median background signal to be below 60 on each microarray; (5) the present call rates to be within 23–37%; (6) the probe signals of two technical duplicates on each microarray to have a Pearson correlation coefficient (r) of at least 0.97; (7) the perfect match/mismatch ratio to be above 5×; and 8) the P95/P05 ratio to be above seven. If at least one of the criteria was violated, the array was rerun.

Microarray data pre-processing

Building prediction models based on tens of thousands of features (gene probes) is an exhaustive and memory consuming process. Moreover, some statistics tools, such as R, have memory limitations [22]. Hence, it is necessary to filter out the features that have little or no contribution to the prediction performance. All gene probes were first filtered using flags to select detected genes. A 'present' flag was defined as a signal detection *p*-value less than 0.1, or average bead number of at least two. Further, a coefficient of variance was calculated by dividing the bead standard deviation by the average expression signal, and probes with values below 0.3 were also flagged as present. Only a probe that had present flags in at least 90% of the samples for all the arrays was kept for further analysis. The data for each of the remaining gene probes was subsequently processed by introducing a missing value (null) for each probe with an average bead number lower than three, or a coefficient of variance above 0.3. The reduced dataset was then *log*2 transformed and any missing values were imputed using the *k*-Nearest Neighbor (*k*-NN)-algorithm with $k = 10$ nearest neighboring values [23].

Next, normalization was applied by subtracting the signal intensity of each probe in each sample by the mean intensity for that sample across all probes (global mean normalization). Finally the data were adjusted for batch effects from chips, using analysis of variance (ANOVA) correction for each individual gene probe [24]. These normalized data were used for all downstream analyses.

Microarray data analysis

For initial discrimination between the two classes of samples, healthy (class 1) and diseased (class 2), we used Canonical Partial Least Squares (CPLS) [25]. Leave-one-out cross-validation (LOOCV) was used to assess the performance of candidate classifiers. The 'pls' package of the freely available R software [26], which can be downloaded from CRAN (cran.r-project.org), was used to perform the analyses. CPLS is an extension of partial least squares (PLS) regression [27], which is a dimension reduction method that enables the use of secondary information about the samples as additional data during model fitting. The secondary data are assumed to be available at the stage of model building, but they are not necessarily available for prediction of future samples. Hence, this information may not be used as direct input to a classifier.

The gene expression data served as predictors for predicting a response matrix which can be split into two sets of variables, a set of primary interest and a set of additional responses. In our setting, the primary response is the health status of the subjects (dummy coded during computations as -1 for class 1, and 1 for class 2), and the secondary responses are the clinical variables and laboratory measurements (Table 3). The secondary responses provide extra input to the CPLS algorithm which may, given that the extra information is relevant, stabilize parameter estimates and improve prediction performance. A new sample was classified as PD based on gene expression if the predicted primary response was larger than zero and as healthy otherwise.

Table 3
Clinical variables included as additional responses in canonical partial least squares modeling.

Order	Variable
1	Site 1
2	Site 2
3	Site 3
4	Gender
5	A260/A280*
6	A260/A230*
7	Concentration*
8	28 S/18 S*
9	RNA integrity number*
10	Age
11	Any chronic disease (other than Parkinson's disease)
12	Cancer
13	Hypertension
14	Diabetes
15	Heart disease
16	Coronary disease
17	Other chronic disease

*Measures of RNA quality.

Microarray data contain up to about 50,000 measurements of gene expression per sample, but usually the sample sizes are small. In addition, they also suffer from class-imbalance. Classifiers built on class-imbalanced data are biased towards the majority class, performing inaccurately on the minority class [24]. In order to prevent this problem from happening, sample balancing was applied in this study by down-weighting the majority class and up-weighting the minority class. The prediction method was fitted using a LOOCV routine. In each test, $n-1$ of the samples are used for training and one sample is for testing (acting as an unseen data set). For each training set consisting of $n-1$ of the samples, sample weights were calculated based on the class-imbalance. The samples from class $j (j = 1, 2)$ were weighted according to:

$$w_j = \frac{0.5 \cdot (n-1)}{n_j}$$

where n_j is the number of samples belonging to class j.

The elimination of features not contributing significantly to the model may improve the accuracy of classification, but also simplify the model and possibly reveal biologically important genes related to PD. Jackknife feature selection was used to select significant gene probes using a p-value of maximum 0.05 as selection criterion. A double leave-one-out cross-validation (DCV) procedure with weighting for class imbalance (Karlsson MK, Lönneborg A, Sæbø S, unpublished data) was carried out in order to avoid overly optimistic performance estimates of the subsets of predictive gene probes. The prediction performance was estimated using the clinical diagnosis as reference, and the accuracy, sensitivity, specificity and the area under the curve (AUC) was calculated.

The partial least squares framework handles multiple and potentially highly correlated features very well, however, it was assumed that it would be favorable to include more gene probes in the final model rather than to minimize the number of features and thus risk losing biological information. We therefore collected significant probes from all segments of a single cross-validation analysis. Any gene probe that was found significant in at least one segment was included in a final feature set from all the training set samples. Gene probes in the final feature set were investigated in terms of networks, pathways and biological function using the Ingenuity Pathway Analysis (IPA) software tool (Ingenuity Systems; Mountain View, CA) [28].

RESULTS

Characteristics of study subjects

This study comprised 154 subjects; 79 subjects with PD and 75 healthy control subjects. Among the 79 PD subjects, 23 were untreated subjects, so called *de novo* PD patients, 14 were recently diagnosed PD patients treated less than 5 years (PD < 5yrs), and 42 were established PD patients treated 5 years or more (PD > 5yrs). Subject demographics are provided in Table 1. Clinical data were available for all subjects. The healthy subjects included 37 men and 38 women, and the PD subjects included 41 men and 38 women. There was no significant group difference for age or gender ($p > 0.01$). PD patient characteristics are provided in Table 2.

Identification of a prediction model for PD in blood

We applied the CPLS method to the microarray data including 17 clinical variables (Table 3) as additional responses. The data set used consisted of all the 154 samples. A LOOCV procedure was applied to validate the prediction performance. Of the 154 subjects, 74 (48.1%) were predicted as healthy and 80 (51.9%) were predicted as PD by the CPLS classifier, and 135/154 subjects (accuracy, $87.7 \pm 5.2\%$) were predicted in agreement with clinical diagnosis. Among the PD subjects, 20 out of 23 (87%) of *de novo* PD subjects were classified in agreement with the clinical diagnosis. The corresponding numbers for the other groups were 13 of

Table 4
Performance characteristics for the data set ($n = 154$) based on LOOCV with 95% confidence interval

Performance characteristics		N	LOOCV (%)
Accuracy		154	87.7 ± 5.2
Sensitivity	All	79	88.6 ± 7.0
	De novo	23	87.0 ± 13.8
	PD < 5yrs	14	92.9 ± 13.5
	PD > 5yrs	42	88.1 ± 9.8
Specificity		75	86.7 ± 7.7
Positive likelihood ratio (PLR)			6.65
Area under the ROC curve (AUC)			0.94

14 (92.9%) for PD < 5yrs patients and 37 of 42 (88.1%) for the PD > 5yrs patients. This gives a total sensitivity of $88.6 \pm 7.0\%$. Among the healthy control subjects 65 of 75 subjects (specificity, $86.7 \pm 7.7\%$) were classified correctly (Table 4). The Positive Likelihood Ratio was 6.65. The distribution of the test scores for patients and controls in the training set is given in Fig. 1A. When plotting the sensitivity versus 1-specificity in a receiver operating characteristics (ROC) curve (Fig. 1B), there was a good separation of the two groups with an AUC of the ROC curve of 0.94.

Applying a DCV procedure together with Jackknife feature selection to the 154 sample set identified a set of 1367 gene probes to be significant at p-level 0.05 in at least one of the single CV segments. On average, 647 gene probes were selected in each segment, with a standard deviation of 33. The performance characteristics of the feature selection and classifier estimation method are presented in Table 5. Of the 79 PD samples in the set, 66 were predicted correctly, while 63 of the 75 healthy control samples were predicted correctly, showing an overall accuracy of 83.8%. Among the PD subjects, 91.3% (21/23) *de novo* PD subjects and 80.4% (45/56) treated PD were classified in agreement with the clinical diagnosis (Fig. 2A). The area under curve (AUC) was 0.89 (Fig. 2B).

Table 5
Performance characteristics for the data set ($n = 154$) based on DCV with jackknife feature selection ($\pm$95% confidence interval)

Performance characteristics		N	DCV (%)
Accuracy		154	83.8 ± 5.8
Sensitivity	All	79	83.5 ± 8.2
	De novo	23	91.3 ± 11.5
	Treated	56	80.4 ± 10.4
Specificity		75	84.0 ± 8.3
Positive likelihood ratio (PLR)			5.22
Area under the ROC curve (AUC)			0.89

Comparison with published blood gene expression studies

Seven studies on gene expression in blood for PD have been published over the last five years [7–13], but only three present classifiers for PD. Scherzer et al. [7]

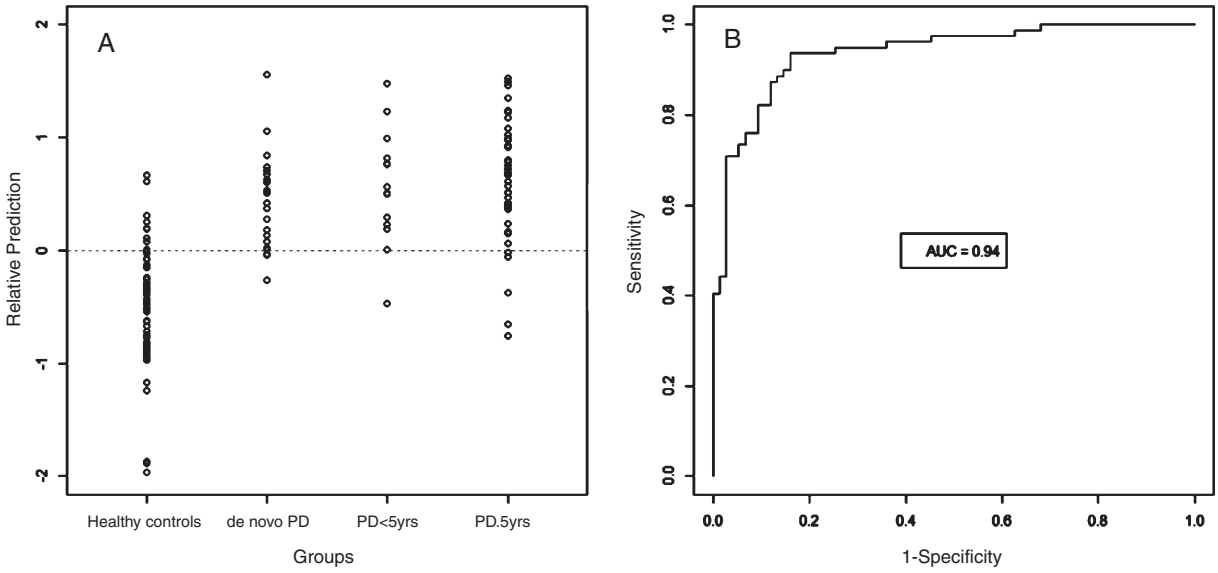

Fig. 1. Prediction performance of the PD classifier based on leave-one-out cross-validation. (A) Classification scores. Among the 79 PD subjects in the data set, 70 were correctly classified, while 65 of the 75 healthy subjects were assigned to the correct class. Twenty of 23 *de novo* PD subjects were correctly classified, 13 of 14 PD < 5yrs subjects were correctly classified, whereas out of 42 PD > 5yrs subjects the classifier predicted the final diagnosis correctly for 37 subjects. A test score > 0 classifies a subject as having PD while a score < 0 classifies a subject as non-PD. (B) ROC curve from data set. Prediction of the 154 subjects based on leave-one-out cross-validation results. Data set gave a classification accuracy of 87.7% and an AUC of 0.94 reflecting a good separation of the two groups.

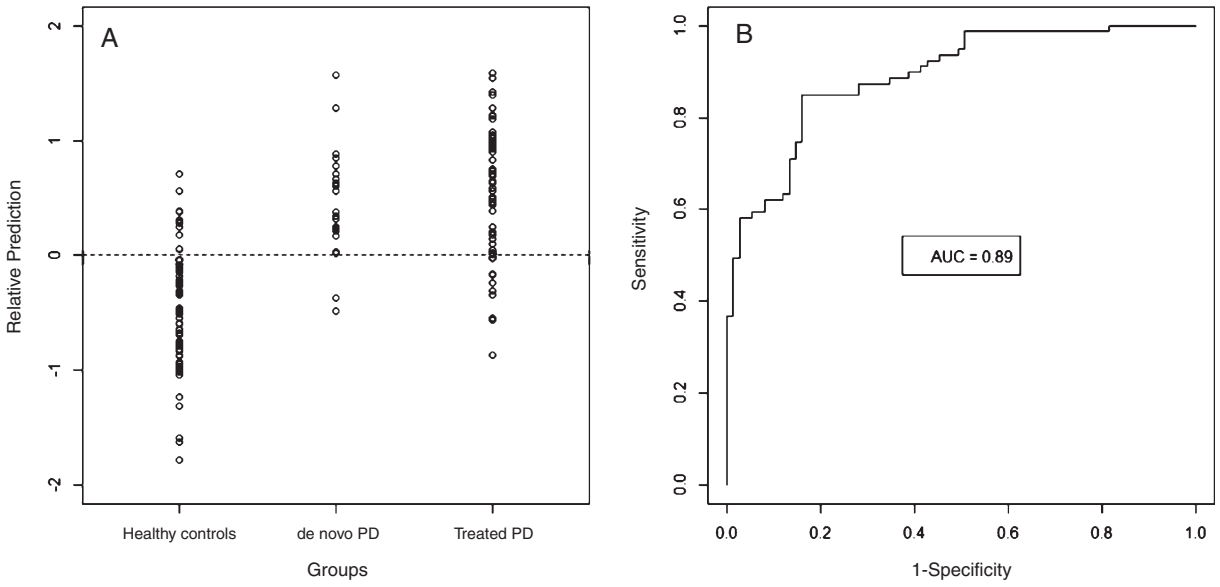

Fig. 2. Prediction performance of the PD classifier based on double leave-one-out cross-validation. (A) Classification scores. Among the 79 PD subjects in the data set, 66 were correctly classified, while 63 of the 75 healthy subjects were assigned to the correct class. Twenty-one of 23 de novo PD subjects were correctly classified, whereas out of 56 treated PD subjects the classifier predicted the final diagnosis correctly for 45 subjects. A test score >0 classifies a subject as having PD while a score < 0 classifies a subject as non-PD. (B) ROC curve from data set. Prediction of the 154 subjects based on double leave-one-out cross-validation results. Data set gave a classification accuracy of 83.8% and an AUC of 0.89 reflecting a good separation of the two groups.

conducted a microarray study and present a set of 22 differentially expressed genes, while Grünblatt et al. [8] and Molochnikov et al. [13] used real-time polymerase chain reaction (RT-PCR) to evaluate twelve and seven genes, respectively. The remaining four studies [9–12] have either not successfully found a discriminating model, or solely perform univariate analysis of differentially expressed genes based on fold change and significance tests. We compared our identified gene list to genes lists presented by Scherzer, Grünblatt and Molochnikov.

Among the set 22 genes identified by Scherzer et al. [7], we found three genes overlapping with our set of genes: the nuclear encoded mitochondrial gene LRPPRC; B-cell CLL/lymphoma 2 (BCL2), which reportedly suppresses apoptosis in a variety of cell systems including neural cells and is linked to the PD-associated gene *Parkin*; and serine/arginine-rich splicing factor 8 (SRSF8/SRP46).

Grünblatt et al. [8] reported a set of four significant genes, but none of these were selected in our analysis. Grünblatt et al. initially evaluated a set of 12 candidate genes selected by postmortem brain profiling, and two of these genes overlap with our gene set: the down-regulated heat shock 70 kDa protein 8 (HSPA8), and the up-regulated ubiquitin-conjugating enzyme E2K (UBC1 homolog, yeast) (UBE2K/HIP2). HSPA8 and HIP2 are similar genes, related to ST13, which has previously been linked to PD by Scherzer et al. Similarly, Molochnikov et al. [13] evaluated a set of seven candidate genes selected from the same postmortem brain profiling study as Grünblatt et al., and presented a five-gene set for differentiating between early PD and controls. Here we found an overlap of two genes: HSPA8 and UBE2K/HIP2, and a third gene, Egl nine homolog 1 (EGLN1) was also found overlapping with the initial set evaluated by Molochnikov et al.

Functional analysis by Ingenuity Pathway Analysis (IPA)

Analysis of differential expression in genes is frequently carried out with the aim to interpret the identified subset of genes in terms of biological functions and pathways. The 1367 selected informative gene probes between the 79 PD patients and 75 healthy control subjects were further explored through pathway and functional analyses using the "Core Analysis" function included in IPA. Out of the 1367 gene probes, a significant number were not annotated ($n = 195$) or had limited biological information and these were removed from the list along with duplicate gene

symbols. One thousand, one hundred and sixteen gene symbols from the total list were recognized by IPA and included in the analysis.

As expected, the functional analysis results identified a number of genes significantly associated to toxicity due to mitochondrial dynamics and function, and oxidative stress ($p < 0.05$). This is in accordance with the established mitochondrial dysfunction in PD [29] and provides further evidence for mitochondrial impairment at the transcriptional level in blood. The significant functions with higher number of genes implicated correspond to 'infectious disease', 'transcription', 'apoptosis', and 'cell death'. Among the most significant canonical pathways were several pathways related to cellular signaling, together with the protein ubiquitination pathway.

Using IPA we identified genes directly interacting between our gene list and for each of the gene lists presented in [7, 8] and [13]. A cluster of 34 genes from our list was found directly interacting with 9 of the 22 genes on Scherzer's list. These 43 genes were associated with gene expression, mainly 'transcription' ($p = 1.2E-12$, Fisher's exact test), but also with 'cellular growth and proliferation' ($p = 1.6E-9$), 'cell death and survival' ($p = 2.8E-8$), and 'oxidative stress' ($p = 1.3E-5$).

Analyzing Grünblatt's 12 gene list, a cluster of 4 genes was found directly interacting with 32 genes from our list. Twelve of these 36 genes were associated with neurological disease, and 11 of which were related to function 'movement disorder' ($p = 4.6E-5$). However, 'cellular growth and proliferation', 'cell death and survival', and 'skeletal and muscular disorders', including 'Parkinson's disease' were also functions significantly associated. The most significant canonical pathway identified was 'protein ubiquitination pathway', and interestingly, prostaglandin J2 emerged as a highly significant upstream regulator with $p = 9.2E-14$ (Fisher's exact test).

Five of the seven genes investigated by Molochnikov et al. [13], were directly interacting with 49 of our 1116 genes. Eleven genes were significantly associated with the protein ubiquitination pathway ($p = 3.7E-10$), and among the most significant functions were 'neurological disease', 'cell death and survival', 'inflammatory response', and several functions related to mitochondrial function and dynamics.

Furthermore, SNCA (alpha-synuclein), the first PD-associated gene identified and believed to have a central part of PD pathology, is one of the identified genes in our list. It has previously been reported to be down-regulated for PD patients [7, 11].

DISCUSSION

Identification of biomarkers for early, preclinical stages of PD is important for the disease treatment and prevention of disease progression to be successful, and towards identifying individuals at risk. This is important since clinical criteria are only applicable late in the disease and here an accurate biomarker for PD has the potential to be a key tool in the process of diagnosis. Further, biomarkers could provide insights into our understanding of the molecular pathogenesis of PD, which in turn, could be used to identify therapeutic targets. Previous studies of gene expression in PD have mainly focused on brain tissue and cerebrospinal fluid (CSF) or animal models [30–32], with limited sample sizes. The present need in a biomarker for PD is for it to be of clinical use. Because brain tissue is obtained at autopsy, it is not a potential candidate for a clinical test. CSF is in direct contact with the brain and spine, but it is more difficult to obtain a spinal fluid sample than a blood sample. Entering the spinal canal with a needle requires expert knowledge and experience to avoid serious complications from the procedure, while blood sampling is a much simpler procedure and can be performed by any trained health professional. More recent genomic studies in PD show an increased interest for gene expression in blood [7–13]. These studies support the hypothesis that changes in gene expression in the brain due to PD can also be found in blood [33], and they also show, together with studies such as [14–16], the potential of developing accurate blood-based tests for major neurological diseases. The present study is the most extensive gene expression profiling for PD yet done in blood. Our analysis of gene expression profiling in peripheral blood from a larger number of samples confirms the findings in previous studies that gene expression in blood may be instrumental in the search for molecular biomarkers for PD.

One of our main findings was that not only did we identify an effective prediction model for PD with 88% accuracy by LOOCV, and 84% accuracy when applying the more robust DCV on a selected subset of the gene probes, but we also found a classifier that predicts *de novo* PD, the earliest clinical stage of the disease, with high accuracy. Since the *de novo* PD patients have not yet been treated for the disease at blood sampling, it strongly suggests that the identified classifiers' predictive ability is not affected by disease treatment. The performance characteristics showed good agreement with clinical diagnosis reaching 87% for LOOCV and 91% for DCV with feature selection. Diagnosis of PD

is usually a straightforward clinical exercise in patients with typical presentation of characteristic symptoms and excellent response to levodopa treatment. Nevertheless, clinicopathological studies have demonstrated difficulties in the diagnosis of the disease in the early stages [34, 35], but also the challenges of differential diagnosis versus atypical parkinsonian disorders, such as progressive supranuclear palsy and multiple systems atrophy, especially early in the disease when signs and symptoms have greater overlap [11]. A blood-based test for the early detection of PD would i) help physicians in difficult cases and potentially identify individuals at risk of the disease; ii) aid clinicians in choosing the best medical treatment at earlier stages; iii) be a useful tool for patient selection when developing novel drugs or other treatments to delay or prevent disease progression.

Functional analysis, using IPA [28], of the selected set of gene probes yielded significant biological functions and canonical pathways previously reported central in PD, such as mitochondrial dysfunction, oxidative stress, neurological disease, cell signaling, and the ubiquitin-proteasome pathway. Functional mitochondria are important for neurotransmission, synaptic maintenance and neuronal survival [29]. Mitochondrial dysfunction is also associated with the generation of oxidative stress, and dysfunctional mitochondria more readily mediate the induction of apoptosis.

Four microarray studies on PD in blood have been conducted over the last five years using mainly the Affymetrix platform. The Illumina platform was selected over others for being quality and cost-effective. We compared our list of identified jack-knife selected genes with some other published gene lists from independent studies on gene expression in blood to determine common findings. Three of the 22 differentially expressed genes presented by Scherzer et al. [7] were present in our gene set. Interestingly, the down-regulated nuclear encoded mitochondrial gene LRPPRC, mentioned in [7] was one of these, and BCL2, which regulates cell death by controlling the mitochondrial membrane permeability, was another. BCL2 has been linked to the PD-associated gene *Parkin*. Of the PD-associated genes, only alpha-synuclein (SNCA) was present in our gene list. SNCA has previously been demonstrated to de down-regulated in PD patients by Scherzer et al. [7] and Soreq et al. [11]. We also found supporting data for a number of the identified differentially expressed genes reported in real-time PCR (RT-PCR) studies [8, 13], which strengthens the results in this study.

The availability of different technologies and platforms for measuring gene expression generates ever more data, and with it, poor reproducibility because different studies analyzing the same clinical outcome report different genes used in the classifiers. A recent study comparing different microarray technologies [36] showed high agreement between data generated by different microarray platforms (correlation = 0.8–0.9), but also that platform differences do exist. Furthermore, the study concluded that sample tissue, signal filtering, and normalization method may also affect the reproducibility. Using small, moderate or large sample sizes for generating gene lists is also an important factor. While we analyzed a relatively balanced set of 154 samples (79 PD and 75 healthy controls), the only two studies of comparable size was the microarray study by Scherzer et al. [7], which compared 66 samples (31 PD and 35 controls), and the RT-PCR study by Grünblatt et al. [8], which compared 139 samples (105 PD and 34 healthy controls).

The gene expression profile identified in this genome-wide microarray study is intended to be further refined and developed into a validated and clinically useful test. Such a test will offer clinicians a convenient blood based tool to supplement the existing diagnostic workup for early diagnosis of PD, as PD is a progressive disorder developing for many years before clinical symptoms become apparent. Considering that the identified gene expression profile predicted *de novo* PD and treated PD patients with equally high accuracy, it is possible that the profile can predict PD at a preclinical stage of the disease. This is yet to be tested.

In this study potential biomarkers for PD were selected by jackknife selection based on the CPLS algorithm. This method is a multivariate approach to variable selection. The identification of a set of relevant, but not redundant, features is central for building prognostic and diagnostic models. Most commonly, individual features are ranked in terms of a quality criterion, such as correlation or *t*-test *p*-values, out of which the top *k* features are selected. However, most feature-ranking methods do not sufficiently account for interactions and correlations between the features, and therefore redundancy is likely to be encountered in the selected features. The results in this paper show that the predictive performance of the CPLS/jackknife method is very good and that the selected set of features are associated with a number of different functions and pathways central in PD. This is an example showing that multivariate approaches may be far more effective than univariate counterparts when it comes to identifying biomarkers for diagnostic purposes. The subtle

differences between sample classes may be much easier to find in the multivariate variable space, than in the limited space spanned by the genes selected on the basis of univariate considerations.

ACKNOWLEDGEMENT

This work is supported by The Norwegian Research Council.

CONFLICT OF INTEREST

M. K. K., P. S. and A. L. are employed by DiaGenic ASA and receive their salaries from the company. P. S. and A. L. are co-founders of DiaGenic ASA and hold substantial stocks in the company. The other authors declare that they have no conflict of interests.

REFERENCES

[1] de Lau LM, Breteler MM (2006) Epidemiology of Parkinson's disease. *Lancet Neurol* **5**, 525-535.

[2] Dorsey ER, Constantinescu R, Thompson JP, Biglan KM, Holloway RG, Kieburtz K, Marshall FJ, Ravina BM, Schifitto G, Siderowf A, Tanner CM (2007) Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology* **68**(5), 384-386.

[3] Hatano T, Kubo SI, Sato S, Hattori N (2009) Pathogenesis of familial Parkinson's disease: New insights based on monogenic forms of Parkinson's disease. *Journal of Neurochemistry* **111**, 1075-1093.

[4] Klein C, Schneider SA, Lang AE (2009) Hereditary parkinsonism: Parkinson disease look-alikes—an algorithm for clinicians to "PARK" genes and beyond. *Mov Disord* **24**, 2042-2058.

[5] Vilariño-Güell C, Wider C, Ross OA, Dächsel JA, Lincoln SJ, Kachergus JM, Soto-Ortolaza AI, Cobb SA, Wilhoite GJ, Bacon JA, Behrouz B, Melrose HL, Hentati E, Puschmann A, Conibear E, Wasserman WW, Aasly JO, Burkhard PR, Djaldetti R, Ghika J, Hentati F, Krygowska-Wajs A, Lynch T, Melamed E, Rajput A, Rajput AH, Solida A, Wu RM, Uitti RJ, Wszolek ZK, Vingerhoets F, Farrer MJ (2011) VPS35 mutations in Parkinson disease. *Am J Hum Genet* **89**(1), 162-167.

[6] Chartier-Harlin MC, Dachsel JC, Vilariño-Güell C, Lincoln SJ, Leprêtre F, Hulihan MM, Kachergus J, Milnerwood AJ, Tapia L, Song MS, Le Rhun E, Mutez E, Larvor L, Duflot A, Vanbesien-Mailliot C, Kreisler A, Ross OA, Nishioka K, Soto-Ortolaza AI, Cobb SA, Melrose HL, Behrouz B, Keeling BH, Bacon JA, Hentati E, Williams L, Yanagiya A, Sonenberg N, Lockhart PJ, Zubair AC, Uitti RJ, Aasly JO, Krygowska-Wajs A, Opala G, Wszolek ZK, Frigerio R, Maraganore DM, Gosal D, Lynch T, Hutchinson M, Bentivoglio AR, Valente EM, Nichols WC, Pankratz N, Foroud T, Gibson RA, Hentati F, Dickson DW, Destée A, and Farrer MJ (2011) Translation initiator EIF4G1 mutations in familial Parkinson disease. *Am J Hum Genet*, **89**(3), 398-406.

[7] Scherzer CR, Eklund AC, Morse LJ, Liao Z, Locascio JJ, Fefer D, Schwarzschild MA, Schlossmacher MG, Hauser MA, Vance JM, Sudarsky LR, Standaert DG, Growdon JH, Jensen RV, Gullans SR (2007) Molecular markers of early Parkinson's disease based on gene expression in blood. *PNAS* **104**, 955-960.

[8] Grünblatt E, Zehetmayer S, Jacob CP, Müller T, Jost WH, Riederer P (2010) Pilot study: Peripheral biomarkers for diagnosing sporadic Parkinson's disease. *J Neural Transm* **117**(12), 1387-1393.

[9] Aguiar PMC, Severino P (2010) Biomarkers in Parkinson disease: Global gene expression analysis in peripheral blood from patients with and without mutations in PARK2 and PARK8. *Einstein* **8**, 293-297.

[10] Shehadeh LA, Yu, K, Wang L, Guevara A, Singer C, Vance J, Papapetropoulos S (2010) SRRM2, a potential blood biomarker revealing high alternative splicing in Parkinson's disease. *PLoS One* **5**(2), e9104.

[11] Soreq L, Israel Z, Bergman H, Soreq H (2008) Advanced microarray analysis highlights modified neuro-immune signaling in nucleated blood cells from Parkinson's disease patients. *J Neuroimmunol* **201-202**, 227-236.

[12] Mutez E, Larvor L, Leprêtre F, Mouroux V, Hamalek D, Kerckaert JP, Pérez-Tur J, Waucquier N, Vanbesien-Mailliot C, Duflot A, Devos D, Defebvre L, Kreisler A, Frigard B, Destée A, Chartier-Harlin MC (2011) Transcriptional profile of Parkinson blood mononuclear cells with LRRK2 mutation. *Neurobiol Aging* **32**(10), 1839-1848.

[13] Molochnikov L, Rabey JM, Dobronevsky E, Bonucelli U, Ceravolo R, Frosini D, Grünblatt E, Riederer P, Jacob C, Aharon-Peretz J, Bashenko Y, Youdim MBH, Mandel SA (2012) A molecular signature in blood identifies early Parkinson's disease. *Molecular Neurodegeneratio* **7**, 26.

[14] Booij BB, Lindahl T, Wetterberg P, Skaane NV, Sæbø S, Feten G, Rye PD, Kristiansen LI, Hagen N, Jensen M, Bårdsen K, Winblad B, Sharma P, Lönneborg A (2011) A gene expression pattern in blood for the early detection of Alzheimer's disease. *J Alzheimers Dis* **23**(1), 109-119.

[15] Rye PD, Booij BB, Grave G, Lindahl T, Kristiansen L, Andersen HM, Horndalsveen PO, Nygaard HA, Naik M, Hoprekstad D, Wetterberg P, Nilsson C, Aarsland D, Sharma P, Lönneborg A (2011) A novel blood test for the early detection of Alzheimer's disease. *J Alzheimers Dis* **23**(1), 121-129.

[16] Fehlbaum-Beurdeley P, Sol O, Désiré L, Touchon J, Dantoine T, Vercelletto M, Gabelle A, Jarrige AC, Haddad R, Lemarié JC, Zhou W, Hampel H, Einstein R, Vellas B (2012) Validation of AclarusDx, a Blood-based transcriptomic signature for the diagnosis of Alzheimer's disease. *J Alzheimers Dis* **32**, 169-181.

[17] Hughes AJ, Daniel SE, Kilford L, Lees AJ (1992) Accuracy of clinical diagnosis of idiopathic Parkinson's disease: A clinico-pathological study of 100 cases. *Journal of Neurology, Neurosurgery, and Psychiatry* **55**(3), 181-184.

[18] Hoehn MM, Yahr MD (1967) Parkinsonism: Onset, progression and mortality. *Neurology* **17**, 427-442.

[19] Fahn S, Elton RL (1987) Members of the UPDRS Development Committee, Unified Parkinson's Disease Rating Scale. In *Recent Development in Parkinson's Disease*, S. Fahn, C.O. Marsden, D.B. Calne, M. Goldstein, eds., Macmillan Health Care Information, Florham Park, NJ, 153-164.

[20] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease (2003) The Unified Parkinson's Disease Rating Scale (UPDRS): Status and recommendations. *Mov Disord* **18**, 738-750.

[21] Folstein MF, Folstein SE, McHugh PR (1975) Mini-Mental State: A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res* **12**, 189-198.

[22] Hornik (2012) "The R FAQ", Version 2.15. http://cran.r-project.org/doc/FAQ/R-FAQ.html

[23] Dasarathy BV (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*, IEEE Computer Society Press, Los Alamitos, Calif.

[24] Tibshirani R, Hastie T, Narasimhan B, Chu G (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS* **99**(10), 6567-6572.

[25] Indahl UG, Liland KH, Næs T (2009) Canonical partial least squares – a unified PLS approach to classification and regression problems. *Journal of Chemometrics* **23**, 495-504.

[26] R Development Core Team (2011) R: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria 2011. http://www. R-project.org, ISBN 3-900051-07-0

[27] Martens H, Næs T (1989) *Multivariate calibration*, Wiley, London.

[28] Ingenuity Pathways Analysis software. http://www.ingenuity.com/

[29] Büeler H (2009) Impaired mitochondrial dynamics and function in the pathogenesis of Parkinson's disease. *Experimental Neurology* **218**(2), 235-246.

[30] Lewis PA, Cookson MR (2011) Gene expression in Parkinson's disease brain. *Brain Research Bulletin* **88**(4), 302-312.

[31] LeWitt P (2012) Recent advances in CSF biomarkers for Parkinson's disease. *Parkinsonism Relat Disord* **18**(1), 49-51.

[32] Blesa J, Phani S, Jackson-Lewis V, Przedborski S (2012) Classic and new animal models of Parkinson's disease. *Journal of Biomedicine and Biotechnology* **2012**, Article ID 845618, pages 10.

[33] Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA (2006) The peripheral blood transcriptome dynamically reflects system wide biology: A potential diagnostic tool. *J Lab Clin Med* **147**(3), 126-132.

[34] Rajput AH, Rozdilsky B, Rajput A (1991) Accuracy of clinical diagnosis in Parkinsonism: A prospective study. *Can J Neurol Sci* **18**(3), 275-278.

[35] Jankovic J, Rajput AH, McDermott MP, Perl DP (2000) The evolution of diagnosis in early Parkinson disease. *Arch Neurol* **57**, 369-372.

[36] Wilder SP, Kaisaki PJ, Argoud K, Salhan A, Ragoussis J, Bihoreau MT, Gauguier D (2009) Comparative analysis of methods for gene transcription profiling data derived from different microarray technologies in rat and mouse models of diabetes. *BMC Genomic* **10**, 63.