Cruz-Monteagudo et al. BMC Medical Genomics (2016) 9:12
DOI 10.1186/s12920-016-0173-x

BMC Medical Genomics

RESEARCH ARTICLE Open Access

Efficient and biologically relevant consensus strategy for Parkinson's disease gene prioritization

Maykel Cruz-Monteagudo[1,2*], Fernanda Borges[1*], Cesar Paz-y-Miño[2], M. Natália D. S. Cordeiro[3], Irene Rebelo[4], Yunierkis Perez-Castillo[5,6], Aliuska Morales Helguera[6], Aminael Sánchez-Rodríguez[7*] and Eduardo Tejera[2]

Abstract

Background: The systemic information enclosed in microarray data encodes relevant clues to overcome the poorly understood combination of genetic and environmental factors in Parkinson's disease (PD), which represents the major obstacle to understand its pathogenesis and to develop disease-modifying therapeutics. While several gene prioritization approaches have been proposed, none dominate over the rest. Instead, hybrid approaches seem to outperform individual approaches.

Methods: A consensus strategy is proposed for PD related gene prioritization from mRNA microarray data based on the combination of three independent prioritization approaches: *Limma*, machine learning, and weighted gene co-expression networks.

Results: The consensus strategy outperformed the individual approaches in terms of statistical significance, overall enrichment and early recognition ability. In addition to a significant biological relevance, the set of 50 genes prioritized exhibited an excellent early recognition ability (6 of the top 10 genes are directly associated with PD). 40 % of the prioritized genes were previously associated with PD including well-known PD related genes such as SLC18A2, TH or DRD2. Eight genes (CCNH, DLK1, PCDH8, SLIT1, DLD, PBX1, INSM1, and BMI1) were found to be significantly associated to biological process affected in PD, representing potentially novel PD biomarkers or therapeutic targets. Additionally, several metrics of standard use in chemoinformatics are proposed to evaluate the early recognition ability of gene prioritization tools.

Conclusions: The proposed consensus strategy represents an efficient and biologically relevant approach for gene prioritization tasks providing a valuable decision-making tool for the study of PD pathogenesis and the development of disease-modifying PD therapeutics.

Keywords: Consensus strategy, Co-expression networks, Early recognition, Gene prioritization, Parkinson's disease

Background

Parkinson's disease (PD) is the second most common neurodegenerative disorder (ND). The present annual cost of health care for patients with PD is estimated to exceed $ 5.6 billion just in the US. With the rapid increase in worldwide life expectancy, the prevalence of PD is expected to double by 2030 [1–3].

Dopamine replacement drugs remains the principal and most effective treatment for PD [4]. However, as the disease progresses, their efficacy diminishes and fails to address the degeneration observed in other brain areas [5–7]. Ultimately, disease-modifying treatments are needed that address both the motor and nonmotor symptoms of PD.

Currently the most important diagnostic marker of PD is limited to the presence of motor disturbances. Unfortunately, due to overlap of symptoms with other neurodegenerative disorders, misdiagnosis is common. Moreover,

* Correspondence: gmailkelcm@yahoo.es; fborges@fc.up.pt; asanchez2@utpl.edu.ec
[1]CIQUP/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal
[7]Departamento de Ciencias Naturales, Universidad Técnica Particular de Loja, Calle París S/N, EC1101608 Loja, Ecuador
Full list of author information is available at the end of the article

motor deficits allowing clinical diagnosis generally appear when 50–60 % of dopaminergic neurons in the *substantia nigra* (SN) are already lost, limiting the effectiveness of potential neuroprotective therapies [8].

In addition to motor symptoms, non-motor symptoms including autonomic dysfunction, depression, olfactory deficit, cognitive disturbances and sleep abnormalities have been related to PD [9]. This mixture of apparently unrelated symptoms and physiological disorders highlight that PD is a multi-causal disorder. Thus, to identify new targets and biomarkers for PD becomes critical for the early diagnosis of this medical condition and for the development of disease-modifying therapies.

In this sense, the systemic picture of gene expression information enclosed in mRNA microarrays experiments encodes relevant clues on the pathogenesis, biomarkers or therapeutics targets for a disease state, but requires of approaches able to unravel it through the accurate prioritization of those disease relevant genes [10]. Several bioinformatics approaches have been reported for this task including those based on differential gene expression [11], gene co-expression networks [12] or machine learning (ML) approaches [13].

Each approach has particular theoretical foundations determining relative advantages and limitations. It is well known that the consensus use of multiple and independent pieces of information increases the reliability of a decision-making process [14]. So, the hybridization of conceptually different approaches can provide prioritization tools with enhanced efficiency [15]. Specifically, such novel hybrid approaches have not been applied yet to PD relevant genes prioritization nor even to neurodegenerative disorders [12]. In this work we propose a consensus strategy for PD relevant genes prioritization based on the integration of several approaches including linear models for microarray data (*Limma*), machine learning, and co-expression networks. Since only a few candidates can usually be considered for further validation experiments, particular emphasis is made in the early recognition ability prioritization tools.

One problem benchmarking the early recognition ability of prioritization approaches in bioinformatics is the lack of statistically sound metrics for this task [16]. Other related areas such as chemoinformatics have standardized procedures to evaluate an analogous problem to gene prioritization, the virtual screening [17]. Here we propose for the first time the use of such early recognition metrics to evaluate the performance of gene prioritization approaches. Hence, besides to identify an enriched set of PD related genes we propose a consensus strategy for gene prioritization with proved enrichment efficiency and biological relevance, as well as a statistically founded approach to evaluate the early recognition ability of gene prioritization tools.

Methods

Microarrays data

Experimental microarray data comparing healthy control (HC) and Parkinson's disease (PD) samples were obtained analyzing the Gene Expression Omnibus (GEO) [18]. Table 1 shows the GEO data sources, references, and sample distribution used in the study. Only studies on *substantia nigra* were considered. So, eight samples collected from *frontal gyrus* were removed from GSE8397.

It is important to highlight that the *substantia nigra* is the region of the brain that shows the greatest loss of dopaminergic neurons in human PD patients. This induce a serious bias that we will term the "dopamine bias". This bias induce a serious risk of overestimation of the enrichment ability of a prioritization strategy based on samples coming from the *substantia nigra*. At the same time, it is also true that dopamine-related process are intrinsically implicated in the pathogenesis of PD. So, we need to check not only which prioritized gene is "dopamine-related", but also whether such gene is associated or not with PD. This critical issue will be considered along all the analysis conducted and properly discussed in the following sections of the manuscript.

Each microarray was processed as follows: public data was extracted and processed using *GEOquery* package in Bioconductor [19]. After individual microarrays analysis, the first step in cross-platform microarray analysis is to combine the different probes. For this task the *entrez gene* was used as identifier in order to obtain the common space across all platforms [20–22]. We mapped the arrays probes of each independent studies to the respective *entrez gene* ID through manual observation and also using the updated manufacturers annotation information (using R-packages: *hgu133a.db*, *hgu133plus2.db* and *hgfocus.db* [23–25]) for all platforms.

Only genes common to all platforms (8477 genes) were used in the subsequent analysis. Genes with more than one probe in individual microarray/studies were combined using the row with the highest mean intensity value applying the *collapseRows* and *intersect* functions implemented in the *WGCNA* package [26, 27]. A second normalization was performed in order to re-scale the intensity and remove cross-platform batch effects using the *Combat* function of the *SVA* package [28]. From the

Table 1 Microarray data details

Code	Platform	Sample	Ref.
GSE20292[a]	GPL96	11(PD); 18(HC)	[45, 53]
GSE7621	GPL570	16(PD); 9(HC)	[46]
GSE20333	GPL201	6(PD); 6(HC)	[97]
GSE8397[b]	GPL96	31(PD); 16(HC)	[47]

[a]Three samples with outlier nature removed after cross-platform normalization
[b]Eight samples collected from *frontal gyrus* removed

initial set of 29 samples in GSE20292 three samples with outlier nature were removed after cross-platform normalization. Finally a subset of 102 samples (59 PD and 43 HC) remained for further analysis.

Differential gene expression analysis

The identification of genes with statistically different expression between HC and PD groups was performed using *lmFit* from *Limma* R-Package [29]. The basic statistic used for significance analysis was the moderated t-statistic after adjustment with the Benjamini and Hochberg's method to control the false discovery rate ("fdr" adjusted *p*-values) [30].

Machine learning analysis

The ML analysis was conducted over a cross-platform normalized microarray data including 8477 common genes for 102 samples. The full data was split up into training and test sets, as part of the validation scheme [31]. Approximately 25 % of the samples were randomly assigned to the "Test Set" by using the *Create a Subset/ Random (Stratified) Sampling* option implemented in STATISTICA 8.0 [32]. Details on the final distribution of the 102 samples can be assessed on Additional file 1: Table S1. Normalized expression values of the 8477 common genes for each of the 102 samples, sample and study identifiers, disease factor (PD or HC), as well as the distribution of training and test samples are provided as supplementary information Additional file 2.

The full vector of 8477 normalized gene expression values was reduced to 500 genes with maximal relevance for the disease factor by means of the minimal redundancy maximal relevance (mRMR) software [33]. Details of the reduced gene set by using the mRMR software are provided in the supplementary information. Then, the reduced vector was subject to an independent process of feature selection relying on eleven different ranking feature selection algorithms implemented on WEKA 3.7.11 [34]. See the full list of attribute evaluators in the supplementary information. Additionally, the reduced vector was subject to a wrapper subset selection using as attribute evaluators only those ML classifiers including a subset feature selection stage implemented on WEKA 3.7.11.

Weighted gene co-expression network construction and analysis

The full set of 8477 common genes was used for weighted genes co-expression network (WGCN) construction in each group using the *WGCNA* package [27]. In this study, we set the β parameter variation to 6, following the scale-free topology criterion proposed by Zhang and Horvath using the *pickSoftThreshold* function in *WGCNA* [35]. Once defined the adjacency matrix for each group (HC and PD), the corresponding co-expression matrices (CoHC and CoPD) were obtained.

Modular analysis

The modules were detected using the *Dynamic Tree Cut* algorithm [36] by using the *cutreeDynamic* function implemented in the *WGCNA* package. Here, the deep split was set to 3, the cutting height to the 99th percentile and the joining heights on the dendograms were set to the maximum. The node connectivity (k) and the node intramodular connectivity (k_{intra}) were calculated for each module as described in [37].

Statistical significance

The gene ontology (GO) and diseases enrichment analysis were performed using *DAVID bioinformatics resource v6.7* [38], exploiting the well know Gene Ontology Annotation (GOA) [39] and Genetic Association (GAD) [40] databases. The ToppCluster tool for the combined enrichment analysis [41] was used to provide network representations of individual and common terms. The statistical significance of the respective enrichment analyses was accessed by using FDR criteria with p-value < 0.05 as cut-off.

The statistical significance of each genes set prioritized as relevant for PD was assessed as proposed by Chen et al. [42, 43]. Detailed information on the application of this test is provided in the supplementary information. Additionally, a bootstrap random sampling experiment was implemented in *R* as proposed by [42, 43] and performed to test the probability of randomly selecting the same number of known PD related genes in the prioritized genes sets. The Wilcoxon signed rank test was used as significance test.

Enrichment and early recognition

Several enrichment metrics have been proposed in the chemoinformatics literature to measure the enrichment ability of a VS protocol [17]. However, despite being bioinformatic's gene prioritization and chemoinformatic's virtual screening essentially the same problem, this type of enrichment analysis has not been applied in bioinformatics. In this work, we use some of the most extended metrics to estimate the enrichment ability of the gene prioritization strategies proposed. The overall enrichment metrics used here include the area under the accumulation curve (*AUAC*); the area under the ROC curve (*ROC*); and the enrichment factor (*EF*) evaluated at the top 1 %/5 %/10 %/20 % of the ranked list. At the same time, the early recognition metrics used were the robust initial enhancement (*RIE*) and the Boltzmann-enhanced discrimination of ROC (*BEDROC*) evaluated at the top 1 %/5 %/10 %/20 % of the ranked list [17]. The calculation of both classic and early recognition

enrichment metrics was conducted by using the perl script *Cresset_VS* [44].

Results and discussion

Limma based gene prioritization

First, the background of 8477 genes provided by the 102 samples of HC and PD patients was processed with *Limma*. The goal here is to identify those single genes significantly differentiated between HC and PD samples and so, potentially associated with PD. This procedure identified a set of 134 genes with an "fdr" adjusted *p-values* < 0.05, each of which was considered to be significantly differentiated on PD patients. Details on this set of genes are reported as supplementary information. The results of the disease enrichment analysis are shown in Table 2. The number of genes associated with PD and included in GAD provides evidence of a statistically significant association of the selected set of genes with PD (p-value = 0.0271).

It is important to note that the GAD database only covers 29 % of the top 134 genes prioritized using an FDR corrected p-value < 0.05 as significance cutoff. Similarly, the OMIM database have only a coverage of just 25 %. Accordingly, the ranking provided by the disease enrichment analysis must be used as reference instead of a exact criterion of the degree of association of the prioritized genes set with the disease. Consequently, the information in Table 2 can be only used to support the statistically significant association between the top 134 genes prioritized by *Limma* and PD.

Table 2 Disease enrichment analysis on the Genetic Association Database of a set of 134 genes prioritized for PD by using *Limma*

GAD Term	*p-Value*	Hits Sample	Total Sample	Hits Background	Total Background
bipolar disorder	0.0030	7	39	96	2459
schizophrenia	0.0034	11	39	249	2459
alcohol abuse	0.0227	4	39	40	2459
Parkinson's disease	0.0271	6	39	112	2459
delinquent behavior violent behavior	0.0307	2	39	2	2459
schizophrenia; opium abuse	0.0307	2	39	2	2459
alcoholism	0.0346	4	39	47	2459
nicotine dependence smoking behavior	0.0457	2	39	3	2459
impulsivity	0.0457	2	39	3	2459
bipolar affective disorder; unipolar affective disorder	0.0457	2	39	3	2459
personality traits	0.0480	3	39	23	2459

Hits Sample: Number of genes selected by *Limma* that are asociated with the disease condition; *Total Sample*: Number of genes selected by *Limma*; *Hits Background*: Number of genes in the background that are asociated with the disease condition; *Total Background*: Number of genes in the background

However, if we use an uncorrected p-value < 0.5 as a significance cutoff instead of the FDR corrected p-value, the set of prioritized genes increases notably to 1016 genes with a non statistically significant association with PD (data not shown). Such a radical change supports the choice in this work to use FDR corrected instead of uncorrected p-values. It could be explained by the wellknwon ability of the FDR correction to minimize the number of false negatives [30] which minimize the lost of PD related genes and consequently, increasing the enrichment of the gene set selected by using this criterion.

The full list of the top 1016 genes prioritized are provided as a suplementary information (see Additional file 5). In this list we can find several genes reported in previous transcriptome analysis based on similar samples [45–51], some using the same micrarray data used in our work. Even so, it is hard to know the real degree of overlapping between our genes and those reported in these works because not every paper reports the full list of significantly differentiated genes. Moreover, in these works several dissimilar processing strategies were applied which impose and additional degree of difficulty on the comparison across these and our study.

If we look for example to the works reported in [47, 48, 51, 52], the degree of overlaping between the genes lists reported is extremely low. Actually, no common genes were found between the four studies and the maximal overlapping between two studies were two common genes (LRRFIP1 and MDH1) between [5] and [6]. Such a minimal degree of overlapping could be atributed to the diversity of tissues, samples or methodological approaches applied on each independent study. However, when the unique set of 243 genes extracted from the combination of the genes sets reported in [47, 48, 51, 52] is compared with our genes prioritized with *Limma*, a significantly higher degree of overlaping is found. Specifically, a 4.92 % of overlapping (50 common genes) is found considering the top 1016 genes (using the uncorrected p-value < 0.05 as a significance cutoff); 8.21 % of overlapping (11 common genes) considering the top 134 genes (using FDR corrected p-value < 0.05); and 6.49 % of overlapping (39 common genes) considering the top 608 genes (using FDR corrected p-value < 0.25). The last top fraction of 608 genes using a cutoff of 0.25 for FDR corrected p-values was also included in the comparison since such a cutoff is widely used in this type of prioritizations [47–50, 53]. One should expect a higher degree of overlapping for larger gene sets. However, as described, the higher degree of overlapping was found in the top 134 genes prioritized by using FDR corrected p-values. Again, the ability of the FDR

correction to minimize the number of false negatives can be the explanation to this unexpected observation.

Other genes known to be associated with PD such as TH, SLC18A2, NR4A2, DDC and SLC6A3 can be found in our *Limma* prioritization. Interestingly, compared with these genes, SNCA exhibited a lower significance. An statistically significant differenced expression of SNCA is considered mandatory for clinical diagnosis of classical PD [8, 48]. In this prioritization we noted this differencial expression (see supplementary information), but just using as a cutoff an adjusted p-value < 0.25, in agreement with previous studies [47–50, 53]. On the other hand, a reduction in dopamine markers as well as the the presence of α-synuclein–positive Lewy bodies in *substantia nigra* are not exclusive of PD [8, 54]. Therefore it is not surprising that the consensus approach prioritized other genes before SNCA.

A different scenario emerges from the GO enrichment analysis of biological proceses. From this analysis, the overall information extracted is that although the set of genes prioritized by *Limma* do not fully match with known genes associated with PD, the biological processes involving these genes are well known to be implicated in the pathogenia of PD. The GO terms, description, and the FDR corrrected p-values corresponding to the top 11 statistically significant biological process identified from the set of 134 genes are provided in Table 3. Details on the full list of biological process associated to this gene set can be accessed in the suplementary information (see Additional file 5).

The information provided in Table 3 clearly reveals an enrichment in dopamine and neurotransmition process. Although the key role of dopamine metabolism in PD is well known [6], the reduction of dopamine synthesis or simply changes in the metabolism of the dopamine are not exclusive of PD. Such effect in other neurodegenerative disorders or even aging has been recently discussed [51]. Additionally, we can not rule out that the enrichment observed in dopamine process could be a possible consequence of a particular degradation in the *substantia nigra* or even a combined factor for neuronal loss in this particularly sensible tissue [48, 50]. Obviously, is not possible to isolate these effects without aditional experimental data. We also found (although with FDR corrected p-values < 0.05) other biological process well stablished in PD such as oxidative fosforilation and energetic metabolism [46–49, 53] (see details in the supplementary information). The lack of statistical significance of these process is obviously a direct consequence of the reduction of the gene set comming from the use FDR corrrected p-values as cutoff. Actually, when the entire set of 1016 genes (using uncorrected p-values) is subject to the same GO enrichment analysis, these processes become significantly more enriched than dopaminergic processes. The details on the GO enrichment analysis are provided as supplementary information (see Additional file 5). This also indicates that even when a bias toward dopamine metabolism exist, additional information relevant to PD is enclosed in the microarray data used. As discussed later, the consensus strategy actually favor the inclusion of such non dopamine related process.

Table 3 GO terms, description, and the FDR corrrected p-values corresponding to the statistically significant biological process identified from 134 genes prioritized by *Limma*

GO terms	Description	p-value (FDR)
GO:0006576	biogenic amine metabolic process	3,3E-04
GO:0042401	biogenic amine biosynthetic process	3,7E-04
GO:0034311	diol metabolic process	8,2E-04
GO:0009712	catechol metabolic process	8,2E-04
GO:0006584	catecholamine metabolic process	8,2E-04
GO:0018958	phenol metabolic process	9,8E-04
GO:0042423	catecholamine biosynthetic process	3,2E-03
GO:0042398	cellular amino acid derivative biosynthetic process	1,4E-02
GO:0042416	dopamine biosynthetic process	2,3E-02
GO:0006575	cellular amino acid derivative metabolic process	2,9E-02
GO:0042417	dopamine metabolic process	4,4E-02

Finally, another important finding to mention is that the transcriptional coactivator PPARGC1A (PGC-1α) was not found to be significantly differenciated in our study, even when it is a master regulator of mitochondrial biogenesis and oxidative metabolism [48, 50]. In this sense, it is important to note that these studies applied different methodologies so to find this gene as not significantly differentiated is a perfectly possible scenario. The fact that only one of the four studies used in this work reported this gene as diferentially expressed support this observation. Finally, even when PPARGC1A was not found in our study, several genes were found to be direct interactors, and biological process directly related with this gene are clearly present in our prioritized genes. It is elaborated further based on the results shown by the functional interaction network of the set of 50 genes finally prioritized.

Machine learning based gene prioritization

For the ML based gene prioritization process, the full vector of 8477 normalized gene expression values was first reduced to 500 genes with maximal relevance for the disease factor (see the full list in the Additional file 3). This set of 500 genes comprises the 91 % of the 134 genes prioritized by *Limma*. This indicates that this initial gene set used as input for feature selection and further ML modeling conserves almost the same information

prioritized by *Limma*. Then, the reduced vector was subject to an independent process of feature selection as previously depicted in Methods section. Once ranked the 500 relevant genes by the respective attribute selection method, each gene is scored according to their mean rank position across the eleven attribute evaluators by applying a desirability function [55]. The corresponding gene relevance score $d(Rank_i)$ is defined as:

$$d(Rank_i) = \frac{Rank_i - 1}{1 - Rank_{max}} \quad 0 \leq \mathrm{d}(Rank_i) \leq 1 \tag{1}$$

Here $Rank_i$ denotes the rank position assigned to the gene i by the attribute evaluator while $Rank_{max}$ is determined by the number of genes to rank and corresponds to the worst possible rank position (500th). Finally, the overall relevance score for a gene i deduced from the consensus ranking analysis $D(Rank_i)$ is computed as the arithmetic mean of the $d(Rank_i)$ values across all the attribute evaluators applied.

Next, the 500 genes previously identified were also subject to a wrapper subset selection as described in Methods section. The relevance of the subset of genes selected is deduced from the accuracy of the respective classifier. So, we only considered as relevant those subset of genes coming from classifiers exhibiting values of accuracy, sensitivity and specificity over 0.6 on training and validation sets. Table 4 provides details of the predictive performance of the thirteen ML classifiers. Considering the classification performance we can assert that based on the set of genes identified by each ML algorithm it is possible to classify the disease status of our microarray samples with a confidence ranging from 75 to 83 % (see Table 4). The sets of genes selected by the respective classifiers are provided in Additional file 1: Table S2.

Again, by applying a desirability function is possible to score the relevance of the respective gene according to the number of valid classifiers including the gene i and so, considering it as relevant. The corresponding gene relevance score based on the consensus classifier analysis $d(Class_i)$ ranges between 0 (only one valid classifier includes the gene) and 1 (the gene is considered relevant by all the valid classifiers) and is defined as:

$$d(Class_i) = \frac{Nrel_i - 1}{N_{Class} - 1} \quad 0 \leq \mathrm{d}(Class_i) \leq 1 \tag{2}$$

Here $Nrel_i$ denotes the number of valid classifiers including the gene i while N_{Class} indicates the number of valid classifiers.

Table 4 Classification performance of the ML classification algorithms used to identify PD relevant sets of genes

ML Classification Algorithm	Training set			LOO CV			5-Fold CV			Test set		
	Acc.	*Se.*	*Sp.*	*Acc.*	*Se.*	*Sp.*	*Acc.*	*Se.*	*Sp.*	*Acc.*	*Se.*	*Sp.*
functions.SimpleLogistic	1.000	1.000	1.000	0.827	0.860	0.781	0.827	0.814	0.844	0.704	0.750	0.636
rules.MODLEM	1.000	1.000	1.000	0.813	0.837	0.781	0.760	0.767	0.750	0.778	0.750	0.818
rules.PART	0.987	0.977	1.000	0.653	0.674	0.625	0.747	0.721	0.781	0.741	0.750	0.727
trees.ADTree	1.000	1.000	1.000	0.853	0.860	0.844	0.787	0.721	0.875	0.741	0.750	0.727
trees.BFTree	0.973	1.000	0.938	0.853	0.884	0.813	0.747	0.744	0.750	0.741	0.750	0.727
trees.FT	1.000	1.000	1.000	0.800	0.837	0.750	0.867	0.884	0.844	0.741	0.813	0.636
trees.LADTree	1.000	1.000	1.000	0.840	0.884	0.781	0.827	0.814	0.844	0.889	0.875	0.909
trees.LMT	1.000	1.000	1.000	0.813	0.860	0.750	0.773	0.767	0.781	0.741	0.813	0.636
trees.SimpleCart	0.973	1.000	0.938	0.827	0.837	0.813	0.747	0.721	0.781	0.741	0.750	0.727
meta.AdaBoostM1	1.000	1.000	1.000	0.840	0.884	0.781	0.880	0.907	0.844	0.926	1.000	0.818
meta.AttributeSelectedClassifier	0.960	0.977	0.938	0.680	0.721	0.625	0.760	0.767	0.750	0.852	0.875	0.818
meta.ClassificationViaRegression	0.960	0.977	0.938	0.813	0.814	0.813	0.733	0.698	0.781	0.815	0.938	0.636
meta.Decorate	1.000	1.000	1.000	0.893	0.860	0.938	0.867	0.837	0.906	0.963	1.000	0.909
AVERAGE	*0.989*	*0.995*	*0.981*	*0.808*	*0.832*	*0.777*	*0.794*	*0.782*	*0.810*	*0.798*	*0.832*	*0.748*

Acc. = accuracy or overall classification rate; *Se.* = sensitivity or true positives rate (% of PD samples correctly classified); *Sp.* = specificity or true negatives rate (% of HC samples correctly classified)

functions.SimpleLogistic: Classifier for building linear logistic regression models [104]; *rules.MODLEM*: Class for building and using a MODLEM algorithm to induce rule set for classification [105]; *rules.PART*: Class for generating a PART decision list [106]; *trees.ADTree*: Class for generating an alternating decision tree [107]; *trees.BFTree*: Class for building a best-first decision tree classifier [108]; *trees.FT*: Classifier for building 'Functional trees', which are classification trees that could have logistic regression functions at the inner nodes and/or leaves [109]; *trees.LADTree*: Class for generating a multi-class alternating decision tree using the LogitBoost strategy [110]; *trees.LMT*: Classifier for building 'logistic model trees', which are classification trees with logistic regression functions at the leaves [104, 111]; *trees.SimpleCart*: Class implementing a classification and regression tree with minimal cost-complexity pruning [112]; *meta.AdaBoostM1*: Metaclassifier class for boosting a nominal class classifier using the Adaboost M1 method [113]; *meta.AttributeSelectedClassifier*: Metaclassifier class where dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/AttributeSelectedClassifier.html; *meta.ClassificationViaRegression*: Metaclassifier class for doing classification using regression methods [114]; *meta.Decorate*: Meta-learner for building diverse ensembles of classifiers by using specially constructed artificial training examples [115, 116]

The final subset of relevant genes proposed by the ML prioritization strategy is determined by 168 unique genes forming the union of the subsets of genes identified by the valid classifiers. Finally, the absolute relevance of each gene ($MLrel_i$) is estimated by considering its respective $D(Rank_i)$ and $d(Class_i)$ scores and quantified as the corresponding arithmetic mean. Details on this set of genes are reported as supplementary information (see Additional file 4).

The final result is a list of 168 unique genes (see Additional file 5) with proved capability of discriminating PD from HC samples, and sorted according to their consensus merit ($MLrel_i$). This ML set was subject to a disease enrichment analysis, providing evidence of a statistically significant association of the selected genes with PD, placing PD 2nd in the list, with p-value = 0.0367. However, none of the biological process involved in this set of genes was statistically significant. It is important to note that ML methods are focused on maximizing the correct classification rate. So, contrary to standard prioritization methods based on gene expression data, the set of genes identified with ML favor the relevance for the disease state instead the gene connectivity information or the biological background. Accordingly, it is unlikely that the final gene list prioritized by ML methods provide statistically significant enrichments of biological processes or pathways.

Table 5 Connectivity, differential expression and machine learning data used as criteria for module prioritization

Healthy Control (HC) Modules										
Module	*n*	<*k*>	<*kintra*>	<*logPD-logHC*>	*nML*	*Merit_ML*	*nLimma*	*Merit_Limma*	*nML-Limma*	*Merit_ML-Limma*
HC_01	123	12.04	1.38	−0.021	3	1.23	1	0.51	1	1.23
HC_02	349	34.57	7.29	−0.061	6	0.87	13	2.36	4	1.73
HC_03	1057	19.04	8.85	0.011	4	0.19	2	0.12	2	0.29
HC_04	169	17.02	2.59	−0.002	1	0.30	0	0.00	0	0.00
HC_05	347	9.23	2.59	0.165	2	0.29	1	0.18	1	0.44
HC_06	74	8.26	0.73	0.005	0	0.00	0	0.00	0	0.00
HC_07	290	14.81	5.19	0.073	4	0.70	6	1.31	1	0.52
HC_08	251	10.94	2.05	0.030	11	2.21	10	2.52	5	3.02
HC_09	2	1.15	0.00	0.022	0	0.00	0	0.00	0	0.00
HC_10	37	15.32	1.48	0.043	1	1.36	0	0.00	0	0.00
HC_11	91	10.95	1.23	0.048	3	1.66	0	0.00	0	0.00
HC_12	61	23.65	3.85	0.028	2	1.65	0	0.00	0	0.00
HC_13	164	10.23	1.79	0.007	3	0.92	1	0.39	1	0.92
HC_14	71	8.33	0.81	−0.001	0	0.00	0	0.00	0	0.00
HC_15	2120	49.53	36.69	−0.062	82	1.95	97	2.89	40	2.86
HC_16	3271	22.06	14.66	−0.064	46	0.71	3	0.06	1	0.05
Parkinson's Disease (PD) Modules										
PD_01	603	286.30	70.52	0.022	6	0.50	1	0.10	1	0.25
PD_02	1437	262.21	150.85	−0.126	69	2.42	103	4.53	42	4.42
PD_03	133	210.12	13.36	0.035	1	0.38	0	0.00	0	0.00
PD_04	161	284.83	22.96	0.089	4	1.25	3	1.18	2	1.88
PD_05	789	231.70	62.45	−0.025	5	0.32	1	0.08	0	0.00
PD_06	468	238.37	38.64	0.132	3	0.32	0	0.00	0	0.00
PD_07	494	316.82	58.43	0.103	24	2.45	19	2.43	8	2.45
PD_08	213	218.15	28.17	−0.033	4	0.95	2	0.59	1	0.71
PD_09	4179	333.39	247.08	−0.047	52	0.63	5	0.08	2	0.07

n: number of genes in the module; <*k*>: average node degree; <k_{intra}>: intra-modular average node degree; <*logPD-logHC*>: module average differential of the log transformed average expression of a gene *i* across PD samples and healthy control samples; n_{ML}: number of genes identified by ML analysis included in the module; n_{Limma}: number of genes identified by *Limma* analysis included in the module; $n_{ML\text{-}Limma}$: number of common genes identified by both ML and *Limma* analyses included in the module; $Merit\_ML = (n_{ML}/168)/(N/8477)$: merit assigned to the module based on n_{ML}, the total number of genes identified by ML analysis (168), *N*, and the total number of background genes (8477); $Merit\_Limma = (n_{Limma}/134)/(N/8477)$: merit assigned to the module based on n_{Limma}, the total number of genes identified by *Limma* analysis (134), *N*, and the total number of background genes (8477); $Merit\_ML\text{-}Limma = (n_{ML\text{-}Limma}/56)/(N/8477)$: merit assigned to the module based on $n_{ML\text{-}Limma}$, the total number of common genes identified by both ML and *Limma* analyses (56), *N*, and the total number of background genes (8477)

Gene co-expression network modules prioritization

Using the *Dynamic Tree Cut* method, 9 and 16 modules were identified in CoHC and CoPD, respectively. Details on the connectivity profile of both co-expression networks are provided in Table 5.

Based on the connectivity information it should be possible to identify those modules enriched with hub genes [56, 57]. In this sense, relatively high values of the modules average node (gene) degree (<*k*>) as well as the average intramodular node degree (<k_{intra}>) can act as relevant indicators of modules potentially enriched with hub genes. From the connectivity information four potentially PD relevant modules are identiffied. PD_07, PD_01, and PD_04 exhibit particularly high values of <*k*> while modules PD_02, and PD_01 show significantly high values of <k_{intra}>. Among these four modules PD_07 stands out as the module with the highest overall connectivity but with barely high intramodular connectivity. On the other hand PD_02 exhibits a significant but inverse profile.

A solid decision can't be made on the only basis of the connectivity information. So, additional information needs to be considered. For this we focused on the differential of the log transformed average expression of a gene *i* across PD samples and HC samples (*logPD-logHC*). The goal here is to identify modules enclosing

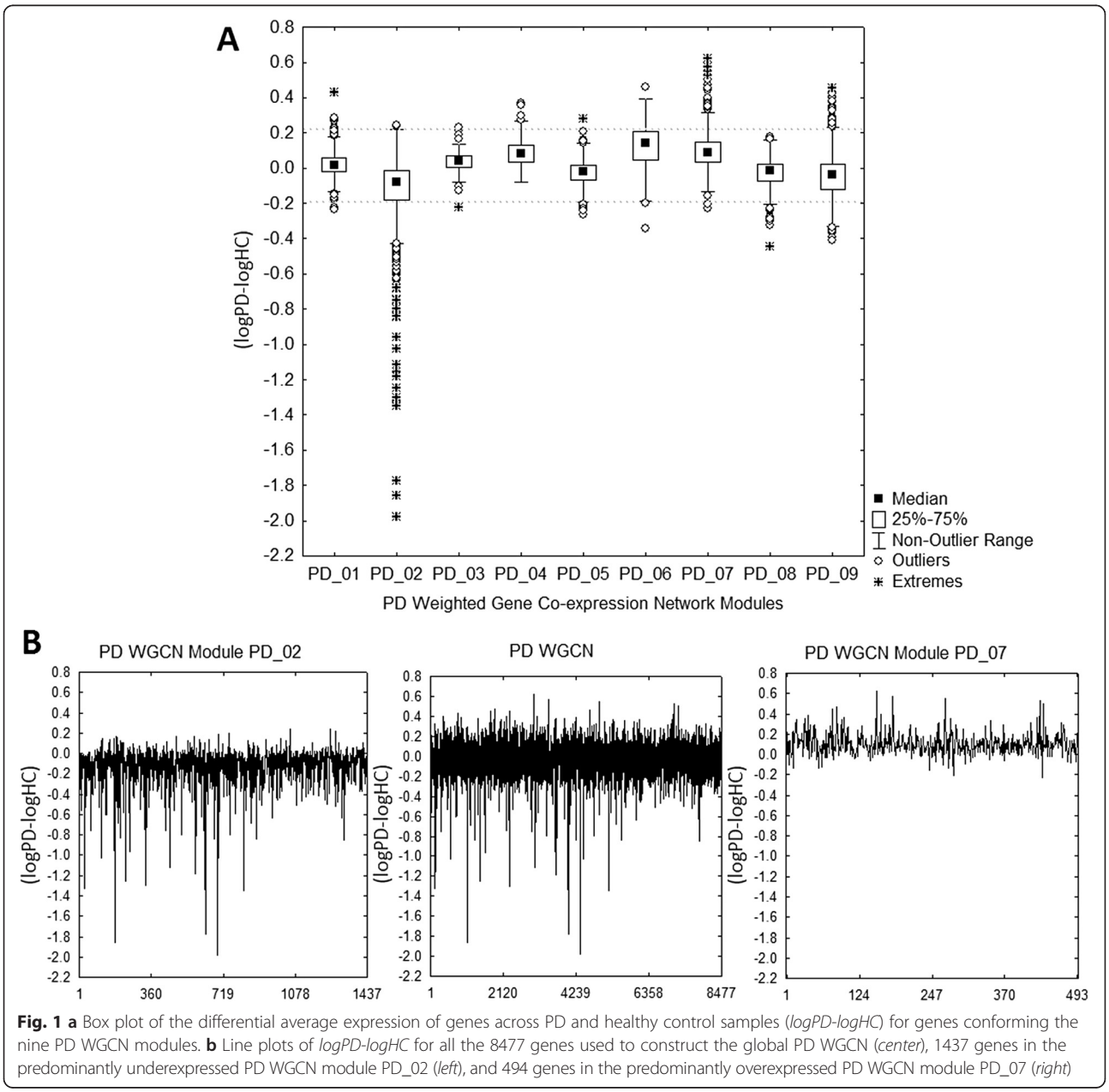

Fig. 1 a Box plot of the differential average expression of genes across PD and healthy control samples (*logPD-logHC*) for genes conforming the nine PD WGCN modules. **b** Line plots of *logPD-logHC* for all the 8477 genes used to construct the global PD WGCN (*center*), 1437 genes in the predominantly underexpressed PD WGCN module PD_02 (*left*), and 494 genes in the predominantly overexpressed PD WGCN module PD_07 (*right*)

genes significantly associated with PD and involved in common biological process that are central in PD [58]. Based on the average *logPD-logHC* value (see Table 5), PD_02 stands out as a significantly underexpressed module while PD_07 toguether with PD_06 are the most overexpressed modules. However, only PD_02 and PD_07 should be selected. From Fig. 1a it is clear that although PD_06 exhibit a slightly higher average *logPD-logHC* value, a significant amount of genes with outlier and extreme behaviour are only present in PD_07. From Fig. 1b it is possible to visually confirm that most of the underexpressed genes in the background (center) belongs to PD_02 (left) while most of the overexpressed genes belongs to PD_07 (right).

It is well known that the consensus use of multiple and independent pieces of information increases the reliability of a decision-making process [14]. So, based on the enrichment potential demonstrated by *Limma* and ML it is feasible to expect a significant confidence gain by incorporating these two independent approaches. From Table 4 can be confirmed the relevance of PD_02 and PD_07 for PD from a ML and/or *Limma* perspective. Here, we use an intuitive measure of the merit of each module based on the number of genes in the module identiffied by each approach. The merit values of ML and/or *Limma* associated to PD_02 and PD_07 outperform from 1.3-fold to 3.8-fold the closest module (PD_04).

Statistical Significance. In order to statistically validate our module prioritization strategy each WGCN PD module was subject to a hypergeometric probability test. Detailed results are provided in Table 6. From this table it is possible to note that only PD_02 is enriched in PD related genes significantly beyond what might be expected by chance (p-value = 0.0034) while PD_07 is in the limits of the statistical significance (p-value = 0.0512). These results support the strategy followed for modules prioritization. Regarding to the inclusion of the module PD_07, as previously mentioned, the GAD database was used just as a common reference framework for comparison purposes. Therefore, the p-values reported must be used as a decision-making criterion instead of a definitive selection/rejection criterion. On the other hand, the biological relevance of this module also grants its inclusion as will be demonstrated in the following section.

Table 6 Hypergeometric test results for the WGCN PD modules based on 319 known PD related genes in GAD and 8477 background genes

Prioritized PD Module	*n*	*m*	*p*-value
PD_01	603	29	0.1014
PD_02	1437	73	0.0034
PD_03	133	6	0.3849
PD_04	161	6	0.5685
PD_05	789	19	0.9897
PD_06	468	15	0.7776
PD_07	494	26	0.0512
PD_08	213	10	0.2813
PD_09	4179	128	0.9997
PD_02 ∪ PD_07	1931	99	0.0003

n: number or genes in the prioritized PD module; *m*: number of known PD related genes in GAD found in the prioritized module; *p*-value: hypergeometric probability of finding by chance *k* or more known PD related genes in a set of *n* prioritized genes

Biological Relevance. The space of biological process covered by the respective PD_02 and PD_07 gene sets was explored by conducting a joined gene ontology (GO) enrichment analysis in order identify commonalities and uniqueness between these two modules. The association between the corresponding biological process and PD were contrasted with the current literature evidence. The full details on the enrichment analysis are provided as supplementary information (see Additional file 5).

From this analysis four processes well known to be associated with PD can be highlighted from the 1437 genes included in the module PD_02: oxidative phosphorylation; intracellular transport; mitochondrion organization; and learning or memory. These results reflect the well-known mitochondrial complex I deficiency [59] (specifically, primary defects in mitochondrial oxidative phosphorylation [60]) leading to oxidative stress, largely associated to PD and their characteristics motor and cognitive impairments [59–63]. In terms of biological processes, the information provided by the genes included in this module and those prioritized by *Limma* is highly consistent. Even so, contrary to *Limma* prioritization, this module do not enrich mainly dopamine metabolism processes but also energetic process. This suggest that the dopamine bias could be actually compensated by combining *Limma* and co-expression analysis.

From the 494 genes involved in PD_07 three processes well known to be associated with PD can be highlighted: protein folding; response to unfolded protein; and response to protein. These processes had being largely reported by other authors [48, 49, 53] and could be associated with the role of α-Synuclein misfolding and aggregation in the pathogenesis of PD [64].

A combined enrichment analysis of the biological process comprised in PD_02 and PD_07 was conducted with aid of the ToppCluster tool [41] (see details in the Additional file 5). The resultant network representation of individual and common biological process for PD_02 and PD_07 is provided in Fig. 2. As can be noted in this figure, both modules share common biological processes including the influence in protein phosphorylation, apoptosis and protein

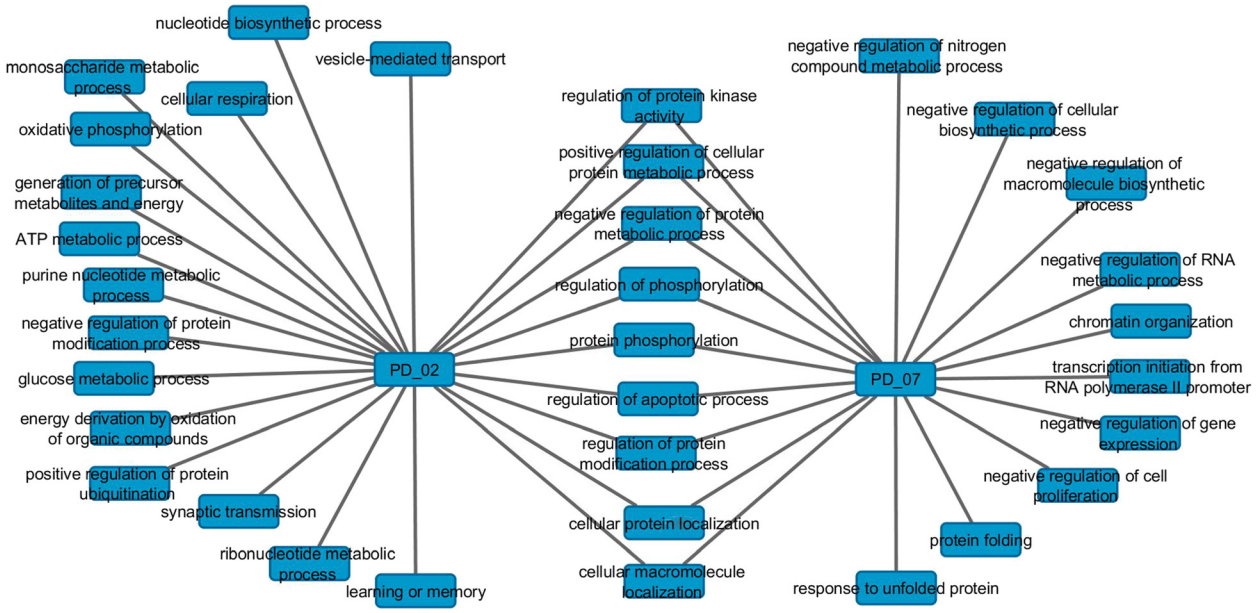

Fig. 2 Representative common and unique biological process covered by modules PD_02 and PD_07

metabolism. Some of these processes, such as oxidative phosphorylation and apoptosis has been extensively reported in PD [46–49, 51, 53], while other process mainly related with post-translational and post-transcriptional modifications have been less explored in PD [48, 49].

For example, SNCA is present in PD_02, however, most of the histones and chaperones are located in PD_07. Specifically the heat shock protein family B (small) member 1 (HSPB1) is included in PD_07. This gene has long been associated with PD [53, 65]. In addition to protein folding this gene is also involved in the apoptosis pathway (11) which is common to both modules. While PD_02 mainly covers energetic and synaptic biological process (oxidative fosforilation, energy metabolism, synaptic transmision and memory), PD_07 is more focused in processes related with folding and transcription regulation origins (protein folding; response to unfolded protein; and response to protein). By considering both modules we are covering not only common biological processes relevant for PD but also other process equally relevant for PD but uniquely covered by the respective module. So, PD_07 not only covers biological process significantly related to PD but also includes some biological process equally significant for PD which are not covered by PD_02.

Consensus gene prioritization strategy

The results obtained in WGCN modules prioritization suggest that the consensus use of several independent sources of information significantly contribute to identify genes sets statistically and biologically relevant to PD. In doing so, all the independent prioritization analyses made (*Limma*, ML, and WGCN analyses) were combined in a consensus gene prioritization strategy. Finding a consensus based on all these tools can provide reliable, statistically significant and biologically relevant genes sets highly enriched with already known and potentially novel PD related genes [14]. The proposed consensus strategy is really simple, but also highly effective as will be demonstrated:

> *Only those genes jointly identified by ML and Limma analysis (common genes) and also present in the biologically relevant WGCN modules PD_02 or PD_07 can be considered as statistically and biologically relevant for PD.*

This consensus strategy based in the common interception of three conceptually different prioritization strategies is actually a highly stringent approach. However; such stringent criteria should provide a desirable balance of enrichment and biological significance of the prioritized gene list.

Our strategy provides a genes list sorted in decreasing order of probability of association with PD by applying fusion rules (*Min-* and *Mean-Rank*) based on *Limma* and ML ranks. That is, genes are first sorted according to the minimum rank assigned by ML and *Limma*, and then by the average of ML and *Limma* ranks.

Following the proposed consensus strategy was prioritized a set of 50 genes sorted in a decreasing order of

relevance for PD. Details on this genes set are provided in Additional file 1: Table S3. As can be noted in the table, 7 out 50 (TP rate = 14 %) genes were found in the set of 319 known PD related genes in GAD. However, after an exhaustive literature search for associations between each of the 50 genes and PD was possible to establish direct associations for 20 genes in this prioritized set (TP rate = 40 %).

Statistical Significance. The statistical validity of the consensus strategy needs to be challenged and compared with the rest of the alternative gene prioritization options. For this, the hypergeometric test, and the random bootstrap sampling were applied to the genes set prioritized by the consensus strategy, the ML and *Limma* analysis (independently and in combination) as well as to the genes set corresponding to PD_02 and PD_07 (independently and in combination). See details in Table 7.

As deduced from the hypergeometric test, not every genes set prioritized can be considered as statistically significant. Although "PD_02 ∪ PD_07" looks like the better option, its significantly higher number of genes compared with "Consensus" hinders its potential for prioritization tasks. Actually, the TP rate of the "Consensus" strategy with only 50 genes is almost three-folds.

Based on the random bootstrap sampling experiment no genes set seems to be randomly enriched with known PD related genes. Again, the consensus strategy stands out for a significantly higher enrichment with known PD related genes compared with the corresponding random enrichment determined in the experiment (*Fold-Enrichment*). The consensus strategy is about four times more enriched in known PD related genes than might be expected by chance, which is almost two-fold compared with "*Limma*", the nearest strategy according to *Fold-Enrichment*.

Enrichment and Early Recognition Ability. Due to the high cost associated to the experimental validation of gene-disease associations and the high number of candidate genes initially considered (thousands), the early recognition ability of a gene prioritization tool should be considered as the ultimate measure of its utility [16]. The estimation of the early recognition ability by statiscally sound metrics is well established in chemoinformatics as part of the validation of virtual screening tools. In this work we propose, for the first time, the use of such metrics for gene prioritization tasks.

From the accumulation curve we can deduce overall enrichment from the area under this curve (*AUAC*) which is defined as:

$$AU\ AC = 1 - \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3}$$

where n is the total number of known disease-related genes in the total background gene set (N) and x_i is the relative rank of the i-th known disease-related gene in the ordered list when their corresponding rank r_i is scaled to N, ($x_i = r_i/N$). So, *AUAC* can be interpreted as the probability that a known disease-related gene, selected from the empirical cumulative distribution function defined by the rank-ordered list, will be ranked before a gene randomly selected from a uniform distribution [17].

The (Receiver Operating Characteristic) ROC curve describes the true positives rate (*TP rate*) for any possible change of the number of selected genes as a function of the false positives rate (*FP rate*) [66]. The area under the ROC curve (*ROC*) can be interpreted as the probability that a known disease-related gene will be ranked earlier than a disease-unrelated gene within a rank-ordered list [17]. The *ROC* metric is defined as:

$$ROC = \frac{AUAC}{R_i} - \frac{R_a}{2R_i} \tag{4}$$

where $R_a = n/N$, and stands for the ratio of known disease-related genes in the dataset, whereas $R_i = N\text{-}n/N$,

Table 7 Statistical validation of the different gene prioritization strategies employed in this work (independently and in combination). Hypergeometric test, random bootstrap sampling experiment and enrichment features of the different gene prioritization strategies

	Hypergeometric Test			Random Bootstrap Sampling (100 Generations)						Enrichment	
	n	*m*	*p-value*	*Mean*	*Median*	*Min.*	*Max.*	*Std. Dev.*	*p-value (W)*	*Fold-Enrichment*	*TP Rate*
Limma	134	10	0.0295	5.0410	5	0	17	2.1852	<0.0001	1.9837	0.0746
ML	168	11	0.0520	6.3211	6	0	22	2.4421	<0.0001	1.7402	0.0655
ML ∪ *Limma*	246	14	0.0805	9.2609	9	0	25	2.9426	<0.0001	1.5117	0.0569
PD_02	1437	73	0.0034	55.4259	55	25	87	6.6392	<0.0001	1.3171	0.0508
PD_07	494	26	0.0512	18.5957	18	2	41	4.1038	<0.0001	1.3982	0.0526
PD_02 ∪ PD_07	1931	99	0.0003	72.6709	73	37	112	7.3516	<0.0001	1.3623	0.0513
Concensus	50	7	0.0025	1.8817	2	0	10	1.3407	<0.0001	3.7200	0.1400

n: number or genes in the prioritized PD module; *m*: number of known PD related genes in GAD found in the prioritized module; *p-value*: hypergeometric probability of finding by chance *k* or more known PD related genes in a set of *n* prioritized genes; *Mean/Median/Min./Max./Std. Dev.*: average/median/minimum/maximum/standard deviation of the number of known PD related genes in GAD included in randomly selected gene sets with the same number of genes as the corresponding set of prioritized genes; *Fold-enrichment*: fold difference between *m* and *Mean* (*Fold-enrichment* = *m*/*Mean*); *TP Rate*: ratio of known PD related genes in *n* (*TP Rate* = *m*/*n*)

and represents the ratio of disease-unrelated genes in the total background gene list.

On the other hand, the enrichment factor (*EF*) takes into account the improvement of the hit rate by a gene prioritization protocol compared to a random selection. This metric has the advantage of answering the question: how enriched in known disease-related genes, the set of n genes that I prioritize will be, compared to the situation where I would just pick the n genes randomly?

$$EF = \frac{m/n}{M/n} \tag{5}$$

where n is the number of genes in the filtered fraction (χ) and m is the number of known disease-related genes retrieved at this fraction, being χ determined by the quotient between n and N ($\chi = n/N$). The maximum value that *EF* can take is $1/\chi$ if $\chi \geq M/N$, N/M if $\chi < M/N$, and the minimum value is zero [17].

However, the "early recognition" ability of a prioritization tool is encoded by just a few enrichment metrics such as the robust initial enhancement (*RIE*) and the Boltzmann-enhanced discrimination of ROC (*BEDROC*) metrics [17]. The *RIE* metric describes how many times the distribution of the ranks for known disease-related genes caused by a prioritization protocol is better than a random rank distribution and is defined as:

$$RIE = \frac{\sum_{i=1}^{n} e^{-\alpha x_i}}{\frac{M}{N}\left(\frac{1-e^{\alpha}}{e^{\alpha/N}-1}\right)} \tag{6}$$

The parameter α is used to assign a higher weight (and so a higher contribution to the *RIE* metric) to known disease-related genes ranked at the beginning than those at the end of the ordered list and can be interpreted as the fraction of the list where the weight is important. Specifically, in this work the *RIE* and also *EF* and *BEDROC* metrics were evaluated at χ = 1 %/5 %/10 %/20 %, which corresponds to values of α = 160.9/32.2/16.1/8, respectively.

However, like *EF*, *RIE* depends on N, R_a and α, which hampers its use in datasets of different size and composition. The other limitation is that unlike *ROC*, *RIE* neither provides a probabilistic interpretation nor a measurement of the enrichment performance above all thresholds [66].

In order to derive a new metric overcoming these limitations Truchon and Bayly proposed the *BEDROC* metric [17].

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}} \tag{7}$$

RIE_{min} and RIE_{max} are obtained when all the known disease-related genes are at the beginning and at the end of the ordered list, respectively.

$$RIE_{min} = \frac{1-e^{\alpha R_a}}{R_a(1-e^{\alpha})} \tag{8}$$

$$RIE_{max} = \frac{1-e^{-\alpha R_a}}{R_a(1-e^{-\alpha})} \tag{9}$$

The *BEDROC* metric is a generalization of the *ROC* metric that includes a decreasing exponential weighting function that adapts it for use in early recognition problems. This metric can be interpreted as the probability that a known disease-related gene ranked by a prioritization protocol will be found before a gene that would come from a hypothetical exponential probability distribution function with parameter α. Thus, *BEDROC* should be understood as a "prioritization usefulness scale" [17].

From the seven prioritization strategies being compared, in Table 8 we estimate and compare the respective overall enrichment and early recognition ability of those four providing a ranked list of genes through all or part of the initial background of 8477 candidate genes.

The ranking provided through the full list of 8477 genes by each strategy is defined by the respective scoring factor employed in the gene prioritization process. Since just a subset of genes is prioritized by each

Table 8 Overall enrichment and early recognition metrics of the four prioritization strategies considered

	Limma	*ML*	*ML-Limma*	*Consensus*
Classic Enrichment Metrics				
AUAC	0.498	0.502	0.495	0.540
ROC	0.498	0.502	0.495	0.541
$EF_{1\%}$	2.855	2.521	2.847	3.164
$EF_{5\%}$	1.449	1.387	1.007	1.512
$EF_{10\%}$	1.038	1.385	0.913	1.510
$EF_{20\%}$	0.975	1.054	1.054	1.321
Early Recognition Metrics				
$RIE_{1\%}$	2.452	2.213	2.403	2.577
$RIE_{5\%}$	1.286	1.438	1.157	1.583
$RIE_{10\%}$	1.089	1.225	1.044	1.400
$RIE_{20\%}$	1.021	1.085	1.008	1.230
$BEDROC_{1\%}$	0.094	0.086	0.094	0.099
$BEDROC_{5\%}$	0.091	0.102	0.083	0.113
$BEDROC_{10\%}$	0.131	0.147	0.125	0.168
$BEDROC_{20\%}$	0.216	0.230	0.214	0.262

strategy, only this fraction is ranked and the remaining genes in the full list of 8477 genes are randomized. The rationale of such a experiment design is to resemble as much as possible the respective prioritization strategy. This randomization strategy is prefered over just to evaluate the respective metrics on the respective prioritized genes set in order to avoid the saturation effect present in small sets with a high ratio of known disease-related genes [17]. The goal here is to evaluate the ability of each prioritization strategy to retrieve the highest fraction possible of those 319 known PD relevant genes in the earliest possible fraction of the respective ordered list. The exact composition of the four respective lists (including ranking and aleatorization rules) is detailed in the supplementary information.

All the values corresponding to *AUAC* and *ROC* metrics provided in Table 8 are close to 0.5, reflecting that the overall enrichment ability of the four prioritization strategies is not better than a random selection. This result, although expected due to the fact that >90 % of the candidate genes are randomized must not be interpreted as a lack of utility of the prioritization strategies. Instead, the real estimation of their utility must focuse on their early recognition ability.

The corresponding values of *EF* at the top fractions studied (1, 5, 10, and 20 %) as well as the early recognition metrics (*RIE*, and *BEDROC*) show that the Consensus strategy compares favorably over the rest of strategies considered, but the difference looks minimal. However, the use of biologically relevant information from PD_02 and PD_07 highlights the advantages of using the Consensus strategy. The comparative overall enrichment and early recognition performance of the four prioritization strategies can be visually confirmed on Fig. 3. As can be noted in Fig. 3b, the enrichment performance of the Consensus strategy clearly outperforms the other three strategies on the top 20 % fraction of the list of 8477 genes considered. The same trend is confirmed in the top 1 % fraction (see Fig. 3c), the most relevant fraction to consider for early recognition assessment [16].

Finally, we evaluated whether each of these prioritization methods ranks a set of known PD genes significantly early than an alternative method. For this, we applied a Wilcoxon signed rank test to compare the ranking provided by the four approaches under study (*Limma*, ML, ML-*Limma* and Consensus) for the 100 % and the top 20 %/10 %/5 % of the 319 PD genes collected from GAD. From this analysis is possible to note that although there is not an evident difference between the early recognition metrics of the four approaches, the consensus strategy ranks the PD genes significantly early than the other three approaches (*Limma*, ML and ML- *Limma*) in all the fractions analyzed [100 % (319 PD Genes in GAD), top 20 % (top 64 PD genes), top 10 % (top 32 PD genes) and top 5 % (top 16 PD genes)]. Only the ranking provided by the consensus strategy for the top 16 PD genes (top 5 %) was not significantly better than the ranking provided by *Limma*. See Table 9 for details.

Biological Relevance. Since the final 50 genes comes from the intersection of the prioritizations made by *Limma*, WGCNA modules, and specially ML, a reduced statistical significance of their biological processes should be expected too, similarly to ML. Most of the top enriched GO terms in the biological process enrichment analysis are associated with PD: dopamine (DA) metabolism [59–63, 67–80]; prepulse inhibition (PPI) [81–86];

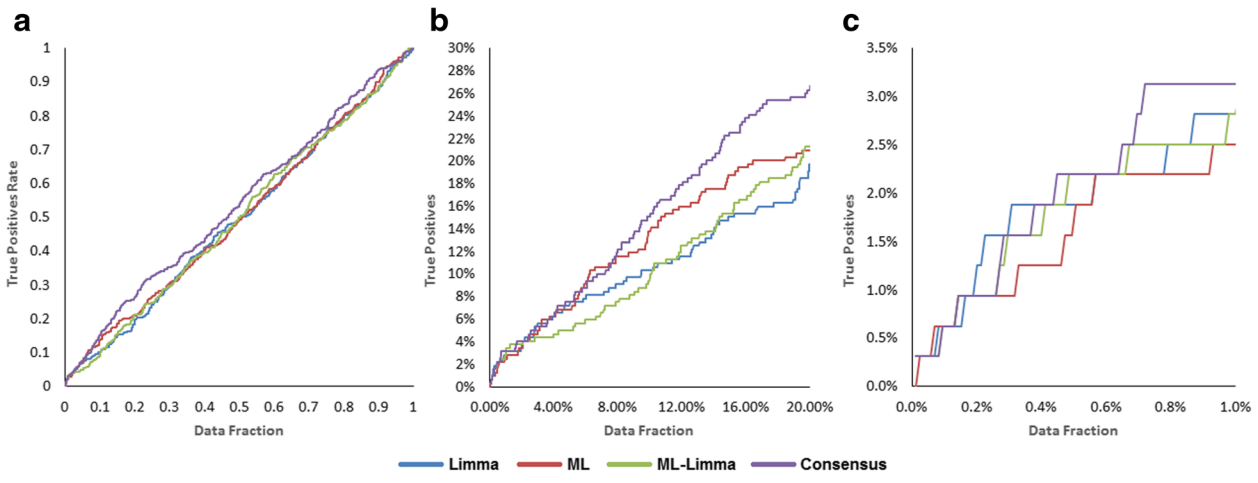

Fig. 3 Accumulation curves of the four prioritization strategies considered. Overall enrichment represented by the accumulation curve for the full set of 8477 background genes for the respective prioritization strategies (**a**). Zoom of the top 20 %/1 % fraction of the ordered list providing information on the early recognition ability of the respective prioritization strategies (**b**/**c**)

Table 9 Results of the Wilcoxon signed rank test conducted to compare the ranking provided by the four approaches under study

319 PD Genes in GAD (100 %)

	Limma	ML	ML-Limma	Consensus
Limma	(−−−)	2.62E-09	2.85E-01	2.79E-62
ML	2.62E-09	(−−−)	1.16E-01	4.36E-47
ML-Limma	2.85E-01	1.16E-01	(−−−)	4.27E-64
Consensus	2.79E-62	4.36E-47	4.27E-64	(−−−)

64 Top Ranked PD Genes in GAD (Top 20 %)

	Limma	ML	ML-Limma	Consensus
Limma	(−−−)	1.84E-05	6.09E-01	6.81E-09
ML	1.84E-05	(−−−)	1.21E-07	1.69E-06
ML-Limma	6.09E-01	1.21E-07	(−−−)	1.24E-12
Consensus	6.81E-09	1.69E-06	1.24E-12	(−−−)

32 Top Ranked PD Genes in GAD (Top 10 %)

	Limma	ML	ML-Limma	Consensus
Limma	(−−−)	7.19E-01	5.23E-04	1.19E-02
ML	7.19E-01	(−−−)	7.25E-02	3.11E-02
ML-Limma	5.23E-04	7.25E-02	(−−−)	1.38E-05
Consensus	1.19E-02	3.11E-02	1.38E-05	(−−−)

16 Top Ranked PD Genes in GAD (Top 5 %)

	Limma	ML	ML-Limma	Consensus
Limma	(−−−)	6.06E-01	4.23E-01	3.02E-01
ML	6.06E-01	(−−−)	3.02E-01	1.95E-03
ML-Limma	4.23E-01	3.02E-01	(−−−)	4.33E-02
Consensus	3.02E-01	1.95E-03	4.33E-02	(−−−)

metal ion transport and pigmentation [87–96]. None of the biological processes is statistically significant by using an FDR adjusted p-value < 0.05 as significance cut-off. See details in the supplementary information. However, from the top ten GO terms only one is directly related with dopamine metabolism pointing to a reduced dopamine bias.

Additionally, an exhaustive literature search was conducted in order to find direct or indirect evidence of the association with PD of each of the 50 genes prioritized. As "direct evidence" we considered scientific publications reporting a relationship (i.e. mutation, expression or knockout) between the gene and PD. As "indirect evidence" we considered scientific publications reporting a theoretical (i.e. system biology) or experimental (i.e. mutation, expression, knockout) evidence of the association of the gene with already known targets or biological processes known to be related with PD pathogenesis.

The microarrays used in our study as raw data correspond to references [45–47, 53, 97]. No result coming from these studies only was used as "evidence". However, studies performing system biology analysis which include also our microarrays were considered because the strategy for data exploration was different and therefore we don't necessarily have to agree in the establishment of genes-diseases association. However, even those studies were considered as "indirect evidence". Any studies carried on in different microarrays and reporting a down/up regulation were considered also but as "indirect evidence".

The literature review conducted evidenced that 20 out of the 50 candidate genes were directly associated with PD (SLC18A2; AGTR1; GBE1; PDCD2; ALDH1A1; SLC6A3; TH; HIST1H2BD; DRD2; EN1; TRIM36; FABP7; PTPRN2; VWA5A; ITPR1; CACNB3; CHORDC1; NDUFA9; RGS4; SNRNP70). Additionally, indirect evidence of association with PD was found for another 8 genes (CCNH; DLK1; PCDH8; SLIT1; BMI1; DLD; PBX1; INSM), which are potentially new therapeutic targets or biomarkers for PD. Details on the direct or indirect literature evidence supporting the association with PD of many of the 50 genes prioritized by our consensus strategy are provided in Table 10.

As previously mentioned, the most relevant feature of the consensus gene prioritization strategy proposed is the early recognition ability evidenced [17]. It is significant that the first 5 genes prioritized (first 10 %) could be confirmed with direct literature evidence. Finally, it is worthy to note that based on the hypergeometric test it is possible to assert that the identification of 20 or more genes out of up to 2402 known PD related genes in a set of 50 prioritized genes is still significantly distant from being a random selection (p-value = 0.049867). That is, considering that an additional set of genes apart of those currently reported in GAD can be relevant for PD but unreported up today, the prioritized list of 50 genes is still statistically significant even in the case that the actual (unknown) set of PD relevant genes would be more than 7-fold (2402) those currently reported in GAD (319).

Considering the above mentioned in addition to the reduced size of the final set of genes prioritized by the consensus strategy we conducted an additional analysis. This analysis was based on the construction of a functional interaction network with the aid of the *Search Tool for the Retrieval of Interacting Genes/Proteins* (STRING) [98, 99] from this final set of 50 genes prioritized with the consensus strategy (actually less because some of these genes don't have reported interaction in our space) and 100 additional interacting genes with a confidence score higher than 0.7. This network was imported into Cytoscape [100] and each gene node was

Table 10 Literature evidence of the association with PD for the 50 genes prioritized with the consensus strategy

Official Gene symbol	Direct Evidence	Indirect Evidence	Description
SLC18A2	1	0	Several studies reported the association between SLC18A2 and PD [117–121]. In humans, the involvement of SLC18A2 in PD pathogenesis is supported by positron emission tomography studies showing significantly lower SLC18A2 densities in the putamen, caudate, and SN of PD patients [122–125]. Its potential as PD biomarker [118] or even as a PD pharmacological target [126] have also been suggested. A method of diagnosing PD comprising a set of differentially expressed genes including SLC18A2 was patented [127].
AGTR1	1	0	AGTR1 have been significantly and consistently downregulated in several PD microarray studies [46, 47, 53, 128, 129]. Additionally, the protective effects on dopaminergic neurons of AGTR1 inhibitors have been well documented [130–136] highlighting the role of AGTR1 as a potential pharmacological target in PD.
GBE1	1	0	GBE1 has been found to be downregulated in gene expression profiling studies of human *substantia nigra pars compacta* from PD patients employing high density microarrays [121, 137]. A method of diagnosing PD comprising a set of differentially expressed genes including GBE1 was patented [127].
PDCD2	1	0	The isoform 1 of PDCD2 was found to be ubiquitinated by parkin and increased in the *substantia nigra* of patients with both autosomal recessive and sporadic PD [138].
ALDH1A1	1	0	ALDH1A1 has been found to be significantly and consistently downregulated in several PD microarray studies [46, 47, 53, 121, 128, 129, 137, 139] highlighting DA metabolism dysfunction resulting in oxidative stress and most probably leading to neuronal cell death. Two methods of diagnosing PD comprising a set of differentially expressed genes including ALDH1A1 were patented [127, 140].
CCNH	0	1	So far, cyclin H (CCNH) has not been directly linked to the pathogenesis of PD. However, the cyclin-dependent kinase 5 (CDK5) was found to act as a mediator of dopaminergic neuron loss in a mouse model of Parkinson's disease [141], pointing the potential role of CCNH as a novel and unexplored PD biomarker.
NRXN3	0	0	No association between NRXN3 and PD was found.
SLC6A3	1	0	A combined analysis of published case–control genetic associations between SLC6A3 and PD involving several ethnicities provided evidences of the role of SLC6A3 as a modest but significant risk factor for PD [142].
DLK1	0	1	No direct associations between DLK1 and PD have been reported. However, through a combined gene expression microarray study in NURR1(−/−) mice DLK1 was identified as novel NURR1 target gene in meso-diencephalic DA neurons [143]. NURR1 (also known as NR4A2) encodes a member of the steroid-thyroid hormone-retinoid receptor superfamily [144]. Mutations in this gene have been associated with disorders related to dopaminergic dysfunction including PD [145–163].
GPR161	0	0	No association between GPR161 and PD was found.
SCN3B	0	0	No association between SCN3B and PD was found.
TH	1	0	TH has been largely associated with PD [164–167].
PCDH8	0	1	No direct association between PCDH8 and PD was found unless a network-based systems biology study utilizing several PD-related microarray gene expression datasets and biomolecular networks [168].
ORC5	0	0	No association between ORC5 and PD was found.
HECA	0	0	No association between HECA and PD was found.
SLIT1	0	1	No direct association between SLIT1 and PD was found. However, the axonal growth inhibition of fetal and embryonic stem cell-derived dopaminergic neurons reported for SLIT1 [169] suggest an indirect association with PD.
BMI1	0	1	Although BMI1 has not been directly associated with PD a previous study demonstrated that it is required in neurons to suppress apoptosis and the induction of a premature aging-like program characterized by reduced antioxidant defenses [170]. These findings provide a molecular mechanism explaining how BMI1 regulates free radical concentrations and reveal the biological impact of BMI1 deficiency on neuronal survival and aging. The activity of BMI1 against mitochondrial ROS may be also relevant to age-associated neurodegenerative diseases where cell death is apparently mediated by oxidative damage, such as in Parkinson disease [171].
QPCT	0	0	No association between QPCT and PD was found.
DLD	0	1	No direct association between DLD and PD was found. However, mice that are deficient in DLD [172] exhibited an increased vulnerability to 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine (MPTP) [173], which have been proposed for use in models of PD [174]. DLD is a critical subunit of key mitochondrial enzyme complexes such as the ketoglutarate dehydrogenase complex (KGDHC) and the pyruvate dehydrogenase complex (PDHC) [175]. Altered energy metabolism, including reductions in KGDHC and PDHC are characteristic of many neurodegenerative disorders including PD [176, 177].
HIST1H2BD	1	0	HIST1H2BD was found to be significantly and differentially expressed in 20 out of the 21 brain regions studied in a multiregional gene expression analysis in postmortem brain coming from 23 control and 22 PD cases [178]. A method of diagnosing PD comprising a set of differentially expressed genes including HIST1H2BD was patented [179].

Table 10 Literature evidence of the association with PD for the 50 genes prioritized with the consensus strategy *(Continued)*

PBX1	0	1	No direct association between PBX1 and PD was found. However, the expression of PBX1 in dopaminergic neurons make it an important player in defining the axonal guidance of the midbrain dopaminergic neurons, with possible implications for the normal physiology of the nigro-striatal system as well as processes related to the degeneration of neurons during the course of PD [180].
SRP72	0	0	No association between SRP72 and PD was found.
DRD2	1	0	DRD2 has been largely associated with PD [181–194].
EN1	1	0	Several studies have reported significant associations between EN1 and PD [195–197].
TRIM36	1	0	TRIM36 has been found to be downregulated in a gene expression profiling study of human *substantia nigra pars compacta* from PD patients employing high density microarrays [137]. A method of diagnosing PD comprising a set of differentially expressed genes including TRIM36 was patented [127].
INSM1	0	1	Although INSM1 has not been directly associated with PD a previous study demonstrated that it is involved on the interrelation of odor and motor changes probably caused by a Mn-induced dopaminergic dysregulation affecting both functions [198]. In this study was found that the rs2871776 G allele, which was associated with the worst effect of Mn on motor coordination, was linked to alteration of a binding site for the transcription factor INSM1. This gene plays an important role in the developing CNS, and especially of olfactory progenitors, as shown in mouse [199] and human [200] embryos. Olfactory impairment is a highly recurrent non-motor dysfunction in PD and is considered an early predictive sign of neurodegeneration [201–203].
MDH2	0	0	No association between MDH2 and PD was found.
CIRBP	0	0	No association between CIRBP and PD was found.
FABP7	1	0	A recent study reported that FABP7 levels were elevated in serum of 35 % of the patients with PD and only in 2 % of the healthy controls, suggesting the role of FABP7 as a potential biomarker for PD [204]. FABP7 was also identified as a promising candidate in a previous quantitative trait loci (QTL) study conducted to identify genes that mediate PPI in mice [205]. This finding was confirmed in a further experiment where FABP7-deficient mice showed decreased PPI. PPI deficiencies is considered a characteristic indicator of schizophrenia [82], but is also deficient in PD patients [206, 207].
PTPRN2	1	0	PTPRN2 has been found to be downregulated in a gene expression profiling study of human *substantia nigra pars compacta* from PD patients employing high density microarrays [137]. A method of diagnosing PD comprising a set of differentially expressed genes including PTPRN2 was patented [127].
PSMG1	0	0	No association between PSMG1 and PD was found.
VWA5A	1	0	VWA5A was associated with PD through a genome-wide genotyping study in PD and neurologically normal controls [208].
ITPR1	1	0	Kitamura et al. [209] reported since 1989 that ITPR1 binding sites were reduced by about 50 % in several brain regions of PD patients (caudate nucleus, putamen, and pallidum) as compared to findings in the age-matched controls, suggesting a probable implication of ITPR1 in PD.
BAI3	0	0	No association between BAI3 and PD was found.
CPT1B	0	0	No association between CPT1B and PD was found.
CACNB3	1	0	The calcium channel subunit b3 (CACNB3), the ATPase type 13A2 (PARK9), and several subunits of Ca^{2+} transporting ATPases (ATP2A3, ATP2B2, and ATP2C1) were downregulated in PD further substantiating the involvement of a deficit in organelle function and of Ca^{2+} sequestering.
ACP2	0	0	No association between ACP2 and PD was found.
CHORDC1	1	0	CHORDC1 was found to be significantly and differentially expressed in 19 out of the 21 brain regions studied in a multiregional gene expression analysis in postmortem brain coming from 23 control and 22 PD cases [178]. A method of diagnosing PD comprising a set of differentially expressed genes including CHORDC1 was patented [179].
SHOC2	0	0	No association between SHOC2 and PD was found.
VBP1	0	0	No association between VBP1 and PD was found.
PPM1B	0	0	No association between PPM1B and PD was found.
YME1L1	0	0	No association between YME1L1 and PD was found.
NDUFA9	1	0	NDUFA9 is included in the KEGG Parkinson's Disease Pathway (http://www.genome.jp/dbget-bin/www_bget?pathway+hsa05012).
TRAPPC2L	0	0	No association between TRAPPC2L and PD was found.
HIST1H2AC	0	0	No association between HIST1H2AC and PD was found.
RGS4	1	0	RGS4 was found to be significantly and differentially expressed in several brain areas of postmortem samples coming from PD patients in comparison to control samples [53]. On the other hand, experiments in mice with reserpine-induced acute DA depletion suggest that RGS4-dependent attenuation of interneuronal autoreceptor

Table 10 Literature evidence of the association with PD for the 50 genes prioritized with the consensus strategy *(Continued)*

			signaling is a major factor in the elevation of striatal acetylcholine release in PD [210]. Lerner and Kreitzer [211] also identified RGS4 as a key link between DA 2/adenosine 2A signaling and endocannabinoid mobilization pathways. In addition, in contrast to wild-type mice, RGS4 deficient mice exhibited normal endocannabinoid-dependent long-term depression after DA depletion and were significantly less impaired in the 6-OHDA model of PD. Taken together, these results suggest that inhibition of RGS4 may be an effective nondopaminergic strategy for treating Parkinson's disease. Finally, RGS4 was recently found to be involved in the generation of abnormal involuntary movements in the unilateral 6-hydroxydopamine (6-OHDA)-lesioned rat model of PD [212].
CRYZL1	0	0	No association between CRYZL1 and PD was found.
RCN2	0	0	No association between RCN2 and PD was found.
SNRNP70	1	0	SNRNP70 was associated with woman affected by PD in an association study of four common polymorphisms in the DJ1 gene and PD involving 416 PD probands and their unaffected siblings matched by gender and closest age [213].
VPS4B	0	0	No association between VPS4B and PD was found.

labeled in order to differentiate those genes in the 50 genes prioritized with the consensus strategy from the 100 additional interacting genes added with STRING. The resultant network representation in provided in the supplementary information (see Additional file 5).

This network includes ubiquitin C (UBC), which appears as a central gene connecting most of the genes included in the network. Although the role of UBC and related genes/proteins in PD through biological process such as protein synthesis, folding and degradation has

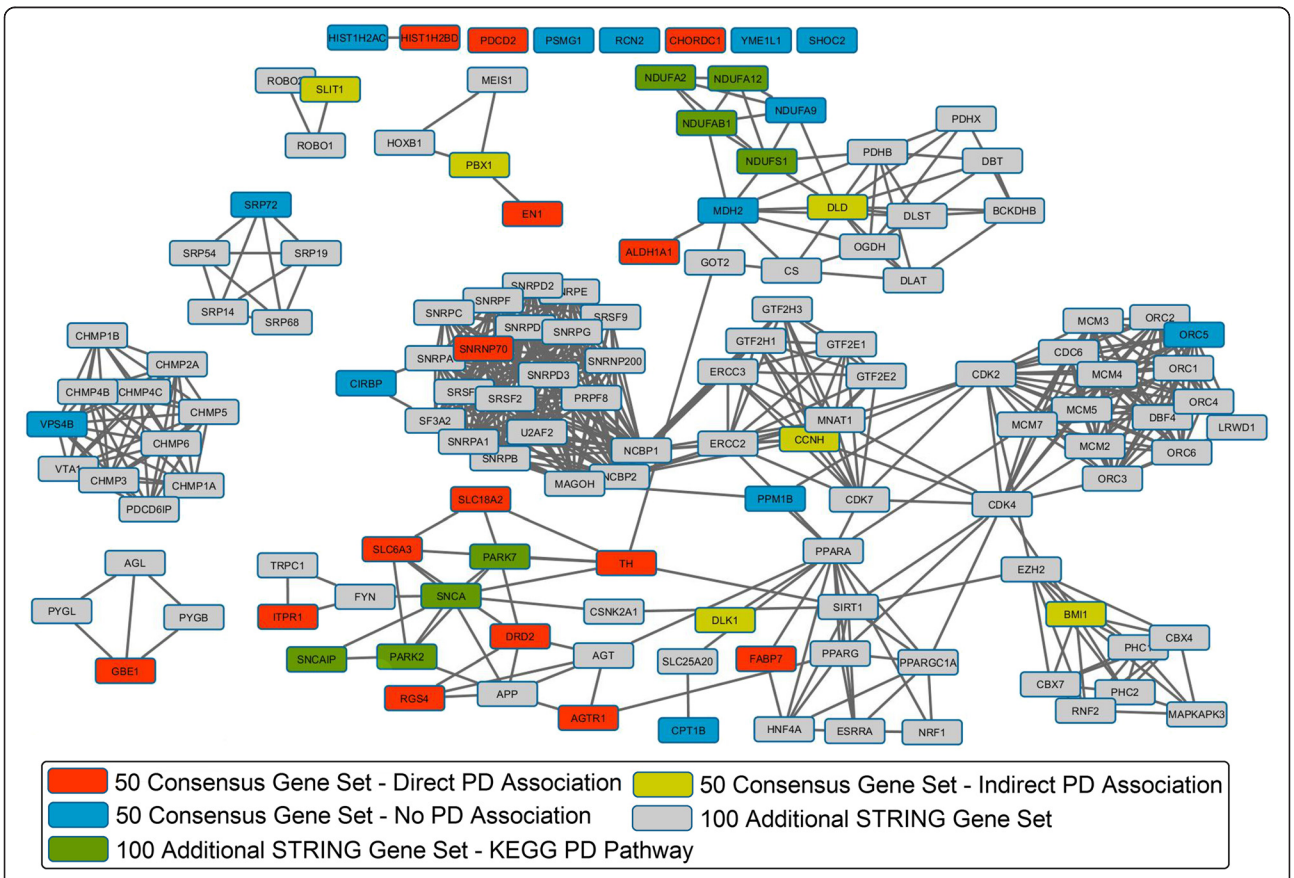

Fig. 4 Functional interaction network of the final set of 50 genes prioritized with the consensus strategy and 100 additional interacting genes. Each gene node was labeled in order to differentiate those genes in the 50 genes prioritized with the consensus strategy from the 100 additional interacting genes (labeled in *gray*). Genes with direct, indirect and no literature evidences of association with PD among the 50 genes prioritized with the consensus strategy were labeled in *red*, *yellow* and *blue*, respectively. Those genes among the 100 additional interacting genes included in the KEGG PD pathway were labeled in *green*

long been established [52, 101, 102], their hub nature in our network could induce a connectivity bias at the time to perform further visual interaction or biological processes enrichment analysis. So UBC was removed from the network previous to conduct the mentioned analysis. Details on the biological processes enrichment analysis are provided in the supplementary information. The functional interaction network after removing UBC is provided in Fig. 4.

From the literature search 22 genes (NRXN3, GPR161, SCN3B, ORC5, HECA, QPCT, SRP72, MDH2, CIRBP, PSMG1, BAI3, CPT1B, ACP2, SHOC2, VBP1, PPM1B, YME1L1, TRAPPC2L, HIST1H2AC, CRYZL1, RCN2, VPS4B) were no associated with PD which challenges the prioritization quality. However; as can be noted in the functional interaction network (see Fig. 5), many of these genes (represented as blue nodes) have a functional connection with important biological processes or genes directly related with PD (represented as red or green nodes). It has to be mentioned that 10 out of these 22 genes (ACP2, BAI3, CRYZL1, GPR161, HECA, NRXN3, QPCT, SCN3B, TRAPPC2L, VPS4B) has no interactions in this space and therefore are not included in this network and that all disconnected clusters and/or nodes in this network are actually connected through UBC gene as can be confirmed in the full network provided in the supplementary information.

An important finding in this network is that even when PPARGC1A was not identified in our study, several genes were found to be direct interactors, and biological process directly related with this gene are clearly present in our prioritized genes. Specifically, can be confirmed that PPARGC1A is connected through short paths with several of the final 50 genes with reported associations with PD (such as TH, AGTR1 and FABP7) or other without current associations with the disease such as PPM1B or CPT1B. On the other hand, the GO enrichment analysis based on this functional interaction network includes several biological process related with the PPARGC1A function. See details in the supplementary information.

The GO enrichment analysis was conducted (based on DAVID) in order to access to significant biological process encoded by the set of genes in this functional interaction network. Contrary to what was expected due

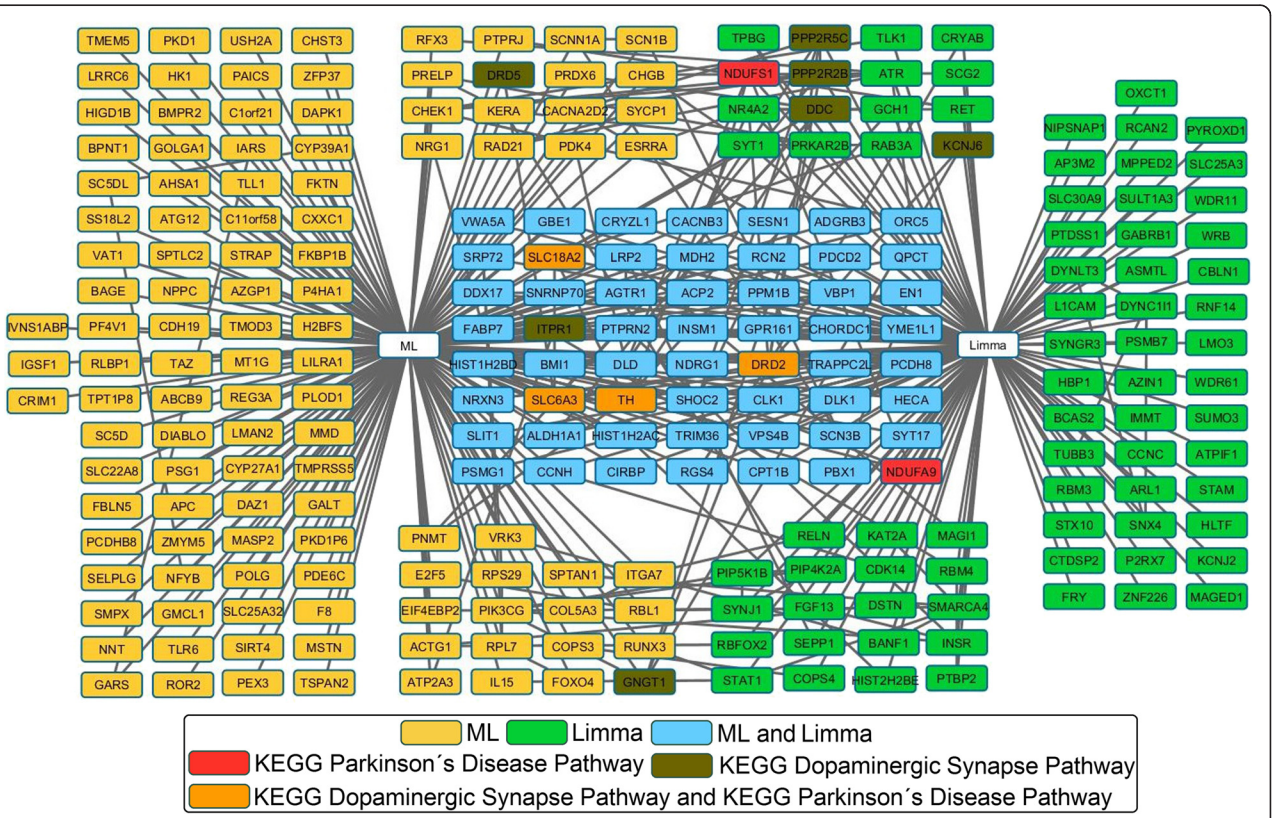

Fig. 5 Functional interaction network comprising gene sets prioritized by *Limma* and ML, respectively. The genes prioritized by ML/*Limma* only are represented by *yellow/green* nodes, while those genes prioritized by both approaches (ML and *Limma*) are represented by *blue* nodes. Genes in the KEGG Dopaminergic Synapse Pathway/KEGG Parkinson's Disease Pathway are represented by *olive/red* nodes, while those genes included in both pathways (Dopaminergic Synapse and Parkinson's Disease) are represented by *orange* nodes

to the risk of the "dopamine bias", from this analysis is clear the highly significant role of RNA splicing [103] (through several mechanisms) and energy metabolism [46–49, 53] compared with the dopamine metabolism process. This last, although statistically significant was placed well below the two former biological process which on the other hand, have been well associated to PD and unrelated to dopamine metabolism. Again, this suggests that the consensus strategy proposed in this work is not affected by the dopamine bias.

Dopamine Bias. As declared from the beginning, the dopamine bias was considered in the discussion of every prioritization method applied. A last experiment was expressly conducted to evaluate this important issue. For this, a functional interaction network was constructed with the aid of STRING from the set of 246 unique genes coming from the union of ML and *Limma* prioritizations (see Fig. 5).

If we look for those genes in the KEGG Dopaminergic Synapse Pathway (129 genes in the DA Pathway) and in the KEGG Parkinson's Disease Pathway (142 genes in the PD Pathway) comprised in the set of 246 unique genes coming from the union of ML and *Limma* prioritizations, it is possible to note that only 4.47 % (11 DA genes out of 246) of this set corresponds to the DA pathway, which indicates an insignificant risk of "dopamine bias" for this set. If we also consider that four out of this eleven DA genes are involved in the PD pathway such risk becomes really insignificant. More importantly, the set of 56 genes shared by ML and *Limma* prioritizations only involves five (DRD2, TH, SLC6A3 and SLC18A2) out of the 129 genes in the KEGG DA pathway. Only one (ITPR1) of these five genes was exclusive of the DA pathway, the other four genes were also included in the KEGG PD pathway. This is a clear indicator of the benefits provided by the integration of conceptually different approaches regarding to avoid the "dopamine bias". All this information can be visually confirmed in the interaction network of genes coming from ML and *Limma* prioritizations provided in Fig. 4. As can be observed in this figure, the ML prioritization is less prone to be affected by the "dopamine bias" which suggest a key role of this approach in reducing such risk.

Finally, only six genes were excluded from the 56 genes from the ML-*Limma* prioritization (CLK1, DDX17, LRP2, NDRG1, SESN1 and SYT17) by concurrently considering the significant PD modules identified in the WGCN analysis (PD_02 and PD_07). Only five out of the 50 prioritized genes were present in the KEGG DA pathway and four out this five dopamine-related genes were included in the KEEG PD pathway. So, from this analysis we can conclude that the consensus strategy proposed in this work is not affected by the "dopamine bias". See details in Table 11.

Table 11 Number of genes in the KEGG DA Pathway, KEGG PD Pathway, and both KEGG DA and PD Pathways in the respective prioritized gene sets

PrioritizationApproach	*N*	*n(%)*			*%*
		DA	*PD*	*DA-PD*	*DA-PD/DA*
ML ∪ *Limma*	246	11(4.47)	6(2.44)	4(1.63)	36.36
ML	168	7(4.17)	5(2.98)	4(2.38)	57.14
Limma	134	9(6.72)	6(4.48)	4(2.99)	44.44
ML ∩ *Limma*	56	5(8.93)	5(8.93)	4(7.14)	80.00
Only-ML	112	2(1.79)	0(0.00)	0(0.00)	0.00
Only-*Limma*	78	4(5.13)	1(1.28)	0(0.00)	0.00
Consensus	50	5(10.00)	5(10.00)	4(8.00)	80.00

N: Number of genes prioritized; *n*: number; *%*: percentage; *DA*: genes in the KEGG Dopaminergic Synapse Pathway; *PD*: genes in the KEGG Parkinson's Disease Pathway; *DA-PD*: genes in the KEGG Dopaminergic Synapse Pathway and the KEGG Parkinson's Disease Pathway

Conclusions

A hydrid gene prioritization approach was applied to PD. Specifically, the set of 50 genes prioritized with the proposed consensus strategy was statistically significant, biologically relevant, highly enriched with know PD related genes and exhibited an excelent early recognition ability. In addition to 20 know PD related genes, eight potentially novel PD biomarkers or therapeutic targets (CCNH, DLK1, PCDH8, SLIT1, DLD, PBX1, INSM1, and BMI1) were identified. Additionally, a statistically rigorous approach of standard use in chemoinformatics was proposed to evaluate the early recognition ability of gene prioritization tools. We also demonstrated that the proper combination of several sources of information is a suitable strategy for module prioritization in co-expression networks analysis. Finally, it is possible to assert that the proposed consensus strategy represents an efficient and biologically relevant approach for gene prioritization tasks, providing a valuable decision-making tool for the study of PD pathogenesis and the development of disease-modifying PD therapeutics.

Additional files

Additional file 1: 1) **Figure S1.** Functional interaction network of the final set of 50 genes prioritized with the consensus strategy and 100 additional interacting genes including UBC. 2) **Table S1.** Samples distribution used for ML analysis. 3) **Table S2.** Sets of PD relevant genes identified by the thirteen ML classification algorithms. 4) **Table S3.** Details on the 50 genes prioritized by means of the proposed consensus strategy. 5) Attribute evaluators used in the consensus ranking analysis. 6) Hypergeometric probability test details. 7) PD related terms in GAD used to identify the set of 513 PD related genes. 8) Composition of the sorted genes lists corresponding to the four prioritization strategies (*Limma*, ML, ML-*Limma*, and Consensus). (DOCX 2196 kb)

Additional file 2: Normalized expression values of the 8477 common genes for each of the 102 samples, sample and study identifiers, disease factor (PD or HC), as well as the distribution of training and test samples. (TXT 10084 kb)

Additional file 3: Details of the reduced gene set by using the mRMR software. (TXT 54 kb)

Additional file 4: Details on the genes sets prioritized by the respective approaches. (TXT 23 kb)

Additional file 5: 1) Results of the *Limma* prioritization for the top 1016 genes with uncorrected *p*-values < 0.05. 2) Results of the gene ontology (biological process) enrichment analysis for the top 134 genes prioritized with *Limma* with FDR corrected *p*-values < 0.05. 3) Results of the gene ontology (biological process) enrichment analysis for the top 1016 genes prioritized with *Limma* with uncorrected *p*-values < 0.05. 4) List of the 168 genes prioritized with machine learning. 5) Results of the gene ontology (biological process) enrichment analysis for the 168 genes prioritized with machine learning. 6) Results of the gene ontology (biological process) enrichment analysis for the 1437 genes included in the co-expression module PD_02. 7) Results of the gene ontology (biological process) enrichment analysis for the 494 genes included in the co-expression module PD_07. 8) Results of the ToppCluster combined enrichment analysis for the co-expression modules PD_02 and PD_07. 9) Results of the gene ontology (biological process) enrichment analysis for the 50 genes prioritized with the consensus strategy and 100 additional interacting genes included in the STRING functional interaction network. (XLSX 493 kb)

Abbreviations

PD: parkinson's disease; ML: machine learning; WGCN: weighted genes co-expression network; GO: gene ontology; *AUAC*: area under the accumulation curve; *ROC*: area under the ROC curve; *BEDROC*: Boltzmann-enhanced discrimination of ROC; *EF*: enrichment factor; *RIE*: robust initial enhancement; TP: true positives; FP: false positives; DA: dopamine; PPI: prepulse inhibition.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ET, AS-R and MC-M conceived, designed and performed the experiments. MC-M wrote the paper. FB and CP-y-M analyzed the data. MNDSC, IR, YP-C and AMH contributed materials and analysis tools. All authors read and approved the final manuscript.

Acknowledgements

Postdoctoral grant [SFRH/BPD/90673/2012] financed by the FCT – Fundação para a Ciência e a Tecnologia, Portugal, co-financed by the European Social Fund. MC-M, ET and CP-y-M acknowledge the financial support from the DITC – Dirección de Investigación y Transferencia de Conocimiento, Universidad de Las Américas – Quito. AS-R acknowledges the financial support from UTPL SmartLand initiative, research program PROY_CCNN_1138.

Author details

[1]CIQUP/Departamento de Química e Bioquímica, Faculdade de Ciências, Universidade do Porto, Porto 4169-007, Portugal. [2]Instituto de Investigaciones Biomédicas (IIB), Universidad de Las Américas, 170513 Quito, Ecuador. [3]REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal. [4]REQUIMTE, Department of Biochemistry, Faculty of Pharmacy, University of Porto, 4050-313 Porto, Portugal. [5]Sección Físico Química y Matemáticas, Departamento de Química, Universidad Técnica Particular de Loja, San Cayetano Alto S/N, EC1101608 Loja, Ecuador. [6]Molecular Simulation and Drug Design Group, Centro de Bioactivos Químicos (CBQ), Central University of Las Villas, Santa Clara 54830, Cuba. [7]Departamento de Ciencias Naturales, Universidad Técnica Particular de Loja, Calle París S/N, EC1101608 Loja, Ecuador.

Received: 3 September 2015 Accepted: 1 March 2016
Published online: 09 March 2016

References

1. Olanow CW, Stern MB, Sethi K. The scientific and clinical basis for the treatment of Parkinson disease (2009). Neurology. 2009;72(21 Suppl 4):S1–136.
2. de Lau LM, Giesbergen PC, de Rijk MC, Hofman A, Koudstaal PJ, Breteler MM. Incidence of parkinsonism and Parkinson disease in a general population: the Rotterdam Study. Neurology. 2004;63(7):1240–4.
3. Dorsey ER, Constantinescu R, Thompson JP, Biglan KM, Holloway RG, Kieburtz K, et al. Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. Neurology. 2007;68(5):384–6.
4. Cotzias GC, Van Woert MH, Schiffer LM. Aromatic amino acids and modification of parkinsonism. New Engl J Med. 1967;276(7):374–9.
5. Braak H, Del Tredici K, Rub U, de Vos RA, Jansen Steur EN, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging. 2003;24(2):197–211.
6. Zarow C, Lyness SA, Mortimer JA, Chui HC. Neuronal loss is greater in the locus coeruleus than nucleus basalis and substantia nigra in Alzheimer and Parkinson diseases. Arch Neurol. 2003;60(3):337–41.
7. Hawkes CH, Del Tredici K, Braak H. A timeline for Parkinson's disease. Parkinsonism Relat Disord. 2010;16(2):79–84.
8. Fearnley JM, Lees AJ. Ageing and Parkinson's disease: substantia nigra regional selectivity. Brain. 1991;114(Pt 5):2283–301.
9. Shulman LM, Taback RL, Bean J, Weiner WJ. Comorbidity of the nonmotor symptoms of Parkinson's disease. Mov Disord. 2001;16(3):507–10. doi:10.1002/mds.1099.
10. Miller RM, Federoff HJ. Microarrays in Parkinson's disease: a systematic approach. NeuroRx : the journal of the American Society for Experimental NeuroTherapeutics. 2006;3(3):319–26. doi:10.1016/j.nurx.2006.05.008.
11. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015. doi:10.1093/nar/gkv007.
12. Gaiteri C, Ding Y, French B, Tseng GC, Sibille E. Beyond modules and hubs: the potential of gene coexpression networks for investigating molecular mechanisms of complex brain disorders. Genes Brain Behav. 2014;13(1):13–24. doi:10.1111/gbb.12106.
13. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9 Suppl 1:S13. doi:10.1186/1471-2164-9-s1-s13.
14. Cisek P. Making decisions through a distributed consensus. Curr Opin Neurobiol. 2012;22(6):927–36. doi:10.1016/j.conb.2012.05.007.
15. Jojic V, Shay T, Sylvia K, Zuk O, Sun X, Kang J, et al. Identification of transcriptional regulators in the mouse immune system. Nat Immunol. 2013;14(6):633–43. doi:10.1038/ni.2587.
16. Doncheva NT, Kacprowski T, Albrecht M. Recent approaches to the prioritization of candidate disease genes. Wiley Interdiscip Rev Syst Biol Med. 2012;4(5):429–42. doi:10.1002/wsbm.1177.
17. Truchon JF, Bayly CI. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. J Chem Inf Model. 2007;47(2):488–508. doi:10.1021/ci600426e.
18. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets—10 years on. Nucleic Acids Res. 2011;39(Database issue):D1005–10. doi:10.1093/nar/gkq1184.
19. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23(14):1846–7. doi:10.1093/bioinformatics/btm254.
20. Rudy J, Valafar F. Empirical comparison of cross-platform normalization methods for gene expression data. BMC Bioinformatics. 2011;12:467. doi:10.1186/1471-2105-12-467.
21. Sirbu A, Ruskin HJ, Crane M. Cross-platform microarray data normalisation for regulatory network inference. PLoS One. 2010;5(11), e13822. doi:10.1371/journal.pone.0013822.
22. Fan X, Shao L, Fang H, Tong W, Cheng Y. Cross-platform comparison of microarray-based multiple-class prediction. PLoS One. 2011;6(1), e16067. doi:10.1371/journal.pone.0016067.
23. Carlson M. hgfocus.db: Affymetrix Human Genome Focus Array annotation data (chip hgfocus). R package version 3.0.0.
24. Carlson M. hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip hgu133plus2). R package version 3.0.0.
25. Carlson M. hgu133a.db: Affymetrix Human Genome U133 Set annotation data (chip hgu133a). R package version 3.0.0.
26. Miller JA, Cai C, Langfelder P, Geschwind DH, Kurian SM, Salomon DR, et al. Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinformatics. 2011;12:322. doi:10.1186/1471-2105-12-322.
27. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559. doi:10.1186/1471-2105-9-559.

28. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27. doi:10.1093/biostatistics/kxj037.
29. Smyth GK. limma: Linear Models for Microarray Data. In: Gentleman R, Carey V, Huber W, Irizarry R, Dudoit S, editors. Bioinformatics and computational biology solutions using R and bioconductor. Statistics for biology and health. New York: Springer; 2005. p. 397–420.
30. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300. doi:10.2307/2346101.
31. Tropsha A. Best practices for QSAR model development, validation, and exploitation. Mol Inf. 2010;29:476–88. doi:10.1002/minf.201000061.
32. StatSoft. STATISTICA. version 8.0 ed. 2007. p. (data analysis software system).
33. Peng H, Long F, Ding C. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Machine Intel. 2005;27(8):1226–38.
34. WEKA. Waikato Environment for Knowledge Analysis (WEKA). 3.7.11 ed. New Zealand: University of Waikato; 2014.
35. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17. doi:10.2202/1544-6115.1128.
36. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. Bioinformatics. 2008;24(5):719–20. doi:10.1093/bioinformatics/btm563.
37. Tejera E, Bernardes J, Rebelo I. Co-expression network analysis and genetic algorithms for gene prioritization in preeclampsia. BMC Med Genet. 2013;6:51. doi:10.1186/1755-8794-6-51.
38. da Huang W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44–57. doi:10.1038/nprot.2008.211.
39. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, et al. The GOA database: Gene Ontology annotation updates for 2015. Nucleic Acids Res. 2015;43(Database issue):D1057–63. doi:10.1093/nar/gku1113.
40. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004;36(5):431–2. doi:10.1038/ng0504-431.
41. Kaimal V, Bardes EE, Tabar SC, Jegga AG, Aronow BJ. ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. Nucleic Acids Res. 2010;38(Web Server issue):W96–102. doi:10.1093/nar/gkq418.
42. Liu X, Liu ZP, Zhao XM, Chen L. Identifying disease genes and module biomarkers by differential interactions. J Am Med Inform Assoc. 2012;19(2):241–8. doi:10.1136/amiajnl-2011-000658.
43. Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. An integrated approach to identify causal network modules of complex diseases with application to colorectal cancer. J Am Med Inform Assoc. 2013;20(4):659–67. doi:10.1136/amiajnl-2012-001168.
44. Mackey MD, Melville JL. Better than random? The chemotype enrichment problem. J Chem Inf Model. 2009;49(5):1154–62. doi:10.1021/ci8003978.
45. Zheng B, Liao Z, Locascio JJ, Lesniak KA, Roderick SS, Watt ML, et al. PGC-1alpha, a potential therapeutic target for early intervention in Parkinson's disease. Sci Transl Med. 2010;2(52):52ra73. doi:10.1126/scitranslmed.3001059.
46. Lesnick TG, Papapetropoulos S, Mash DC, Ffrench-Mullen J, Shehadeh L, de Andrade M, et al. A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. PLoS Genet. 2007;3(6), e98. doi:10.1371/journal.pgen.0030098.
47. Moran LB, Duke DC, Deprez M, Dexter DT, Pearce RK, Graeber MB. Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. Neurogenetics. 2006;7(1):1–11. doi:10.1007/s10048-005-0020-2.
48. Diao H, Li X, Hu S, Liu Y. Gene expression profiling combined with bioinformatics analysis identify biomarkers for Parkinson disease. PLoS One. 2012;7(12), e52319. doi:10.1371/journal.pone.0052319.
49. Zhang B, Xia C, Lin Q, Huang J. Identification of key pathways and transcription factors related to Parkinson disease in genome wide. Mol Biol Rep. 2012;39(12):10881–7. doi:10.1007/s11033-012-1985-1.
50. Su X, Chu Y, Kordower JH, Li B, Cao H, Huang L, et al. PGC-1alpha Promoter Methylation in Parkinson's Disease. PLoS One. 2015;10(8), e0134087. doi:10.1371/journal.pone.0134087.
51. Glaab E, Schneider R. Comparative pathway and network analysis of brain transcriptome changes during adult aging and in Parkinson's disease. Neurobiol Dis. 2015;74:1–13. doi:10.1016/j.nbd.2014.11.002.
52. Lehman NL. The ubiquitin proteasome system in neuropathology. Acta Neuropathol. 2009;118(3):329–47. doi:10.1007/s00401-009-0560-x.
53. Zhang Y, James M, Middleton FA, Davis RL. Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am J Med Genet B Neuropsychiatr Genet. 2005;137B(1):5–16. doi:10.1002/ajmg.b.30195.
54. DelleDonne A, Klos KJ, Fujishiro H, Ahmed Z, Parisi JE, Josephs KA, et al. Incidental Lewy body disease and preclinical Parkinson disease. Arch Neurol. 2008;65(8):1074–80. doi:10.1001/archneur.65.8.1074.
55. Derringer G, Suich R. Simultaneous optimization of several response variables. J Quality Technol. 1980;12(4):214–9.
56. Watts DJ, Strogatz SH. Collective dynamics of 'small-world' networks. Nature. 1998;393(6684):440–2. doi:10.1038/30918.
57. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001;411(6833):41–2. doi:10.1038/35075138.
58. Wolfe CJ, Kohane IS, Butte AJ. Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks. BMC Bioinformatics. 2005;6:227. doi:10.1186/1471-2105-6-227.
59. Schapira AH, Cooper JM, Dexter D, Clark JB, Jenner P, Marsden CD. Mitochondrial complex I deficiency in Parkinson's disease. J Neurochem. 1990;54(3):823–7.
60. Shoffner JM, Watts RL, Juncos JL, Torroni A, Wallace DC. Mitochondrial oxidative phosphorylation defects in Parkinson's disease. Ann Neurol. 1991;30(3):332–9. doi:10.1002/ana.410300304.
61. Perfeito R, Cunha-Oliveira T, Rego AC. Revisiting oxidative stress and mitochondrial dysfunction in the pathogenesis of Parkinson disease–resemblance to the effect of amphetamine drugs of abuse. Free Radic Biol Med. 2012;53(9):1791–806. doi:10.1016/j.freeradbiomed.2012.08.569.
62. Subramaniam SR, Chesselet MF. Mitochondrial dysfunction and oxidative stress in Parkinson's disease. Prog Neurobiol. 2013;106–107:17–32. doi:10.1016/j.pneurobio.2013.04.004.
63. Hauser DN, Hastings TG. Mitochondrial dysfunction and oxidative stress in Parkinson's disease and monogenic parkinsonism. Neurobiol Dis. 2013;51:35–42. doi:10.1016/j.nbd.2012.10.011.
64. Breydo L, Wu JW, Uversky VN. α-Synuclein misfolding and Parkinson's disease. Biochim Biophys Acta (BBA) - Mol Basis Dis. 2012;1822(2):261–85. http://dx.doi.org/10.1016/j.bbadis.2011.10.002.
65. Chen CM, Lee LC, Soong BW, Fung HC, Hsu WC, Lin PY, et al. SCA17 repeat expansion: mildly expanded CAG/CAA repeat alleles in neurological disorders and the functional implications. Clin Chim Acta. 2010;411(5–6):375–80. doi:10.1016/j.cca.2009.12.002.
66. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection–what can we learn from earlier mistakes? J Comput-Aided Mol Design. 2008;22(3–4):213–28. doi:10.1007/s10822-007-9163-6.
67. Bisaglia M, Filograna R, Beltramini M, Bubacco L. Are dopamine derivatives implicated in the pathogenesis of Parkinson's disease? Ageing Res Rev. 2014;13:107–14. doi:10.1016/j.arr.2013.12.009.
68. Rees JN, Florang VR, Eckert LL, Doorn JA. Protein reactivity of 3,4-dihydroxyphenylacetaldehyde, a toxic dopamine metabolite, is dependent on both the aldehyde and the catechol. Chem Res Toxicol. 2009;22(7):1256–63. doi:10.1021/tx9000557.
69. Marchitti SA, Deitrich RA, Vasiliou V. Neurotoxicity and metabolism of the catecholamine-derived 3,4-dihydroxyphenylacetaldehyde and 3,4-dihydroxyphenylglycolaldehyde: the role of aldehyde dehydrogenase. Pharmacol Rev. 2007;59(2):125–50. doi:10.1124/pr.59.2.1.
70. Wey MC, Fernandez E, Martinez PA, Sullivan P, Goldstein DS, Strong R. Neurodegeneration and motor dysfunction in mice lacking cytosolic and mitochondrial aldehyde dehydrogenases: implications for Parkinson's disease. PLoS One. 2012;7(2), e31522. doi:10.1371/journal.pone.0031522.
71. Jinsmaa Y, Florang VR, Rees JN, Mexas LM, Eckert LL, Allen EM, et al. Dopamine-derived biological reactive intermediates and protein modifications: implications for Parkinson's disease. Chem Biol Interact. 2011;192(1–2):118–21. doi:10.1016/j.cbi.2011.01.006.

72. Eisenhofer G, Kopin IJ, Goldstein DS. Catecholamine metabolism: a contemporary view with implications for physiology and medicine. Pharmacol Rev. 2004;56(3):331–49. doi:10.1124/pr.56.3.1.
73. Fornstedt B, Rosengren E, Carlsson A. Occurrence and distribution of 5-S-cysteinyl derivatives of dopamine, dopa and dopac in the brains of eight mammalian species. Neuropharmacology. 1986;25(4):451–4.
74. Klegeris A, Korkina LG, Greenfield SA. Autoxidation of dopamine: a comparison of luminescent and spectrophotometric detection in basic solutions. Free Radic Biol Med. 1995;18(2):215–22.
75. Hastings TG, Lewis DA, Zigmond MJ. Role of oxidation in the neurotoxic effects of intrastriatal dopamine injections. Proc Natl Acad Sci U S A. 1996; 93(5):1956–61.
76. Leroy E, Boyer R, Auburger G, Leube B, Ulm G, Mezey E, et al. The ubiquitin pathway in Parkinson's disease. Nature. 1998;395(6701):451–2. doi:10.1038/26652.
77. Kitada T, Asakawa S, Hattori N, Matsumine H, Yamamura Y, Minoshima S, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. Nature. 1998;392(6676):605–8. doi:10.1038/33416.
78. Cuervo AM, Stefanis L, Fredenburg R, Lansbury PT, Sulzer D. Impaired degradation of mutant alpha-synuclein by chaperone-mediated autophagy. Science. 2004;305(5688):1292–5. doi:10.1126/science.1101738.
79. Hirsch EC, Hunot S, Hartmann A. Neuroinflammatory processes in Parkinson's disease. Parkinsonism Relat Disord. 2005;11 Suppl 1:S9–15. doi: 10.1016/j.parkreldis.2004.10.013.
80. Beal MF. Mitochondria, oxidative damage, and inflammation in Parkinson's disease. Ann N Y Acad Sci. 2003;991:120–31.
81. Braff DL, Grillon C, Geyer MA. Gating and habituation of the startle reflex in schizophrenic patients. Arch Gen Psychiatry. 1992;49(3):206–15.
82. Swerdlow NR, Geyer MA. Using an animal model of deficient sensorimotor gating to study the pathophysiology and new treatments of schizophrenia. Schizophr Bull. 1998;24(2):285–301.
83. Sanchez-Ramos JR, Ortoll R, Paulson GW. Visual hallucinations associated with Parkinson disease. Arch Neurol. 1996;53(12):1265–8.
84. Fenelon G, Mahieux F, Huon R, Ziegler M. Hallucinations in Parkinson's disease: prevalence, phenomenology and risk factors. Brain. 2000;123(Pt 4):733–45.
85. Fernandez HH. Nonmotor complications of Parkinson disease. Cleve Clin J Med. 2012;79 Suppl 2:S14–8. doi:10.3949/ccjm.79.s2a.03.
86. Alobaidi H, Pall H. The role of dopamine replacement on the behavioural phenotype of Parkinson's disease. Behav Neurol. 2013;26(4):225–35. doi:10.3233/ben-2012-120265.
87. Gorell JM, Johnson CC, Rybicki BA, Peterson EL, Kortsha GX, Brown GG, et al. Occupational exposures to metals as risk factors for Parkinson's disease. Neurology. 1997;48(3):650–8.
88. Rybicki BA, Johnson CC, Uman J, Gorell JM. Parkinson's disease mortality and the industrial use of heavy metals in Michigan. Mov Disord. 1993;8(1): 87–92. doi:10.1002/mds.870080116.
89. Zayed J, Campanella G, Panisset JC, Ducic S, Andre P, Masson H, et al. Parkinson disease and environmental factors. Rev Epidemiol Sante Publique. 1990;38(2):159–60.
90. Zayed J, Ducic S, Campanella G, Panisset JC, Andre P, Masson H, et al. Environmental factors in the etiology of Parkinson's disease. Can J Neurol Sci. 1990;17(3):286–91.
91. Altschuler E. Aluminum-containing antacids as a cause of idiopathic Parkinson's disease. Med Hypotheses. 1999;53(1):22–3. doi:10.1054/mehy.1997.0701.
92. Gorell JM, Johnson CC, Rybicki BA, Peterson EL, Kortsha GX, Brown GG, et al. Occupational exposure to manganese, copper, lead, iron, mercury and zinc and the risk of Parkinson's disease. Neurotoxicology. 1999;20(2–3):239–47.
93. Gorell JM, Rybicki BA, Cole Johnson C, Peterson EL. Occupational metal exposures and the risk of Parkinson's disease. Neuroepidemiology. 1999; 18(6):303–8.
94. Dexter DT, Carayon A, Javoy-Agid F, Agid Y, Wells FR, Daniel SE, et al. Alterations in the levels of iron, ferritin and other trace metals in Parkinson's disease and other neurodegenerative diseases affecting the basal ganglia. Brain. 1991;114(Pt 4):1953–75.
95. Riederer P, Sofic E, Rausch WD, Schmidt B, Reynolds GP, Jellinger K, et al. Transition metals, ferritin, glutathione, and ascorbic acid in parkinsonian brains. J Neurochem. 1989;52(2):515–20.
96. Hirsch EC, Brandel JP, Galle P, Javoy-Agid F, Agid Y. Iron and aluminum increase in the substantia nigra of patients with Parkinson's disease: an X-ray microanalysis. J Neurochem. 1991;56(2):446–51.
97. Jacob J. Gene expression profiling of parkinsonian substantia nigra (Expression profiling by array, Homo sapiens). 2013.
98. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447–52. doi:10.1093/nar/gku1003.
99. Snel B, Lehmann G, Bork P, Huynen MA. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. 2000;28(18):3442–4.
100. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504. doi:10.1101/gr.1239303.
101. Lim KL, Tan JM. Role of the ubiquitin proteasome system in Parkinson's disease. BMC Biochem. 2007;8 Suppl 1:S13. doi:10.1186/1471-2091-8-s1-s13.
102. Kim HJ, Kim HJ, Jeong JE, Baek JY, Jeong J, Kim S, et al. N-terminal truncated UCH-L1 prevents Parkinson's disease associated damage. PLoS One. 2014;9(6), e99654. doi:10.1371/journal.pone.0099654.
103. La Cognata V, D'Agata V, Cavalcanti F, Cavallaro S. Splicing: is there an alternative contribution to Parkinson's disease? Neurogenetics. 2015;16(4): 245–63. doi:10.1007/s10048-015-0449-x.
104. Landwehr N, Hall M, Frank E. Speeding up logistic model tree induction. In: Sumner M, Frank E, Hall M, editors. 9th European Conference on Principles and Practice of Knowledge Discovery in Databases; October 3–7; Porto, Portugal. 2005. p. 675–83.
105. Stefanowski J. The rough set based rule induction technique for classification problems. 6th European Congress on Intelligent Techniques and Soft Computing; September 7–10; Aachen, Germany. 1998. p. 109–13.
106. Frank E, Witten IH. Generating accurate rule sets without global optimization. Fifteenth International Conference on Machine Learning; July 24–26; Madison, Wisconsin, USA. 1998. p. 144–51.
107. Freund Y, Mason L. The alternating decision tree learning algorithm. Sixteenth International Conference on Machine Learning; June 27–30; Bled, Slovenia. 1999. p. 124–33.
108. Friedman J, Hastie T, Tibshirani R. Additive logistic regression : a statistical view of boosting. Ann Stat. 2000;28(2):337–407.
109. Gama J. Functional trees. Mach Learn. 2004;55(3):219–50.
110. Holmes G, Pfahringer B, Kirkby R, Frank E, Hall M. Multiclass alternating decision trees. 12th European Conference on Machine Learning; September 5–7; Freiburg, Germany. 2001. p. 161–72.
111. Landwehr N, Hall M, Frank E. Logistic model trees. Mach Learn. 2005;95(1–2): 161–205.
112. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Boca Ratón, FL: Chapman and Hall/CRC Press; 1984.
113. Freund Y, Schapire RE. Experiments with a new boosting algorithm. Thirteenth International Conference on Machine Learning; July 3–6; Bari, Italy. 1996. p. 148–56.
114. Frank E, Wang Y, Inglis S, Holmes G, Witten I. Using model trees for classification. Mach Learn. 1998;32(1):63–76. doi:10.1023/a:1007421302149.
115. Melville P, Mooney RJ. Creating diversity in ensembles using artificial data. Information Fusion. 2005;6(1):99–111. http://dx.doi.org/10.1016/j.inffus.2004.04.001.
116. Melville P, Mooney RJ. Constructing diverse classifier ensembles using artificial training examples. Eighteenth International Joint Conference on Artificial Intelligence; August 9–15; Acapulco, Mexico. 2003. p. 505–10.
117. Brighina L, Riva C, Bertola F, Saracchi E, Fermi S, Goldwurm S, et al. Analysis of vesicular monoamine transporter 2 polymorphisms in Parkinson's disease. Neurobiol Aging. 2013;34(6):1712. e9-13. doi:10.1016/j.neurobiolaging.2012.12.020.
118. Sala G, Brighina L, Saracchi E, Fermi S, Riva C, Carrozza V, et al. Vesicular monoamine transporter 2 mRNA levels are reduced in platelets from patients with Parkinson's disease. J Neural Transm. 2010;117(9):1093–8. doi: 10.1007/s00702-010-0446-z.
119. Alter SP, Lenzi GM, Bernstein AI, Miller GW. Vesicular integrity in Parkinson's disease. Curr Neurol Neurosci Rep. 2013;13(7):362. doi:10.1007/s11910-013-0362-3.
120. Rilstone JJ, Alkhater RA, Minassian BA. Brain dopamine-serotonin vesicular transport disease and its treatment. New Engl J Med. 2013;368(6):543–50. doi:10.1056/NEJMoa1207281.
121. Mandel S, Grunblatt E, Riederer P, Amariglio N, Jacob-Hirsch J, Rechavi G, et al. Gene expression profiling of sporadic Parkinson's disease substantia nigra

pars compacta reveals impairment of ubiquitin-proteasome subunits, SKP1A, aldehyde dehydrogenase, and chaperone HSC-70. Ann N Y Acad Sci. 2005; 1053:356–75. doi:10.1196/annals.1344.031.
122. Okamura N, Villemagne VL, Drago J, Pejoska S, Dhamija RK, Mulligan RS, et al. In vivo measurement of vesicular monoamine transporter type 2 density in Parkinson disease with (18)F-AV-133. J Nucl Med. 2010;51(2): 223–8. doi:10.2967/jnumed.109.070094.
123. Martin WR, Wieler M, Stoessl AJ, Schulzer M. Dihydrotetrabenazine positron emission tomography imaging in early, untreated Parkinson's disease. Ann Neurol. 2008;63(3):388–94. doi:10.1002/ana.21320.
124. Lee CS, Samii A, Sossi V, Ruth TJ, Schulzer M, Holden JE, et al. In vivo positron emission tomographic evidence for compensatory changes in presynaptic dopaminergic nerve terminals in Parkinson's disease. Ann Neurol. 2000;47(4):493–503.
125. Bohnen NI, Albin RL, Koeppe RA, Wernette KA, Kilbourn MR, Minoshima S, et al. Positron emission tomography of monoaminergic vesicular binding in aging and Parkinson disease. J Cereb Blood Flow Metab. 2006;26(9):1198–212. doi:10.1038/sj.jcbfm.9600276.
126. Bernstein AI, Stout KA, Miller GW. The vesicular monoamine transporter 2: an underexplored pharmacological target. Neurochem Int. 2014;73:89–97. doi:10.1016/j.neuint.2013.12.003.
127. Grunblatt E, Mandel SA, Riederer P, Youdim MBH. Diagnostic test for parkinson's disease. Google Patents. 2005.
128. Cantuti-Castelvetri I, Keller-McGandy C, Bouzou B, Asteris G, Clark TW, Frosch MP, et al. Effects of gender on nigral gene expression and parkinson disease. Neurobiol Dis. 2007;26(3):606–14. doi:10.1016/j.nbd.2007.02.009.
129. Bossers K, Meerhoff G, Balesar R, van Dongen JW, Kruse CG, Swaab DF, et al. Analysis of gene expression in Parkinson's disease: possible involvement of neurotrophic support and axon guidance in dopaminergic cell death. Brain Pathol. 2009;19(1):91–107. doi:10.1111/j.1750-3639.2008.00171.x.
130. Sonsalla PK, Coleman C, Wong LY, Harris SL, Richardson JR, Gadad BS, et al. The angiotensin converting enzyme inhibitor captopril protects nigrostriatal dopamine neurons in animal models of parkinsonism. Exp Neurol. 2013;250: 376–83. doi:10.1016/j.expneurol.2013.10.014.
131. Rey P, Lopez-Real A, Sanchez-Iglesias S, Munoz A, Soto-Otero R, Labandeira-Garcia JL. Angiotensin type-1-receptor antagonists reduce 6-hydroxydopamine toxicity for dopaminergic neurons. Neurobiol Aging. 2007;28(4):555–67. doi:10.1016/j.neurobiolaging.2006.02.018.
132. Munoz A, Rey P, Guerra MJ, Mendez-Alvarez E, Soto-Otero R, Labandeira-Garcia JL. Reduction of dopaminergic degeneration and oxidative stress by inhibition of angiotensin converting enzyme in a MPTP model of parkinsonism. Neuropharmacology. 2006;51(1):112–20. doi:10.1016/j.neuropharm.2006.03.004.
133. Kurosaki R, Muramatsu Y, Kato H, Watanabe Y, Imai Y, Itoyama Y, et al. Effect of angiotensin-converting enzyme inhibitor perindopril on interneurons in MPTP-treated mice. Eur Neuropsychopharmacol. 2005;15(1):57–67. doi:10.1016/j.euroneuro.2004.05.007.
134. Lopez-Real A, Rey P, Soto-Otero R, Mendez-Alvarez E, Labandeira-Garcia JL. Angiotensin-converting enzyme inhibition reduces oxidative stress and protects dopaminergic neurons in a 6-hydroxydopamine rat model of Parkinsonism. J Neurosci Res. 2005;81(6):865–73. doi:10.1002/jnr.20598.
135. Saavedra JM, Sanchez-Lemus E, Benicky J. Blockade of brain angiotensin II AT1 receptors ameliorates stress, anxiety, brain inflammation and ischemia: therapeutic implications. Psychoneuroendocrinology. 2011;36(1):1–18. doi:10.1016/j.psyneuen.2010.10.001.
136. Joglar B, Rodriguez-Pallares J, Rodriguez-Perez AI, Rey P, Guerra MJ, Labandeira-Garcia JL. The inflammatory response in the MPTP model of Parkinson's disease is mediated by brain angiotensin: relevance to progression of the disease. J Neurochem. 2009;109(2):656–69. doi:10.1111/j.1471-4159.2009.05999.x.
137. Grunblatt E, Mandel S, Jacob-Hirsch J, Zeligson S, Amariglo N, Rechavi G, et al. Gene expression profiling of parkinsonian substantia nigra pars compacta; alterations in ubiquitin-proteasome, heat shock protein, iron and oxidative stress regulated proteins, cell adhesion/cellular matrix and vesicle trafficking genes. J Neural Transm. 2004;111(12):1543–73. doi:10.1007/s00702-004-0212-1.
138. Fukae J, Sato S, Shiba K, Sato K, Mori H, Sharp PA, et al. Programmed cell death-2 isoform1 is ubiquitinated by parkin and increased in the substantia nigra of patients with autosomal recessive Parkinson's disease. FEBS Lett. 2009;583(3):521–5. doi:10.1016/j.febslet.2008.12.055.
139. Durrenberger PF, Grunblatt E, Fernando FS, Monoranu CM, Evans J, Riederer P, et al. Inflammatory pathways in Parkinson's disease; a BNE microarray study. Parkinsons Dis. 2012;2012:214714. doi:10.1155/2012/214714.
140. Mandel SA, Youdim MBH, Riederer P, Grunblatt E, Rabey JM, Molochnikov L. Peripheral blood gene markers for early diagnosis of parkinson's disease. Google Patents. 2013.
141. Smith PD, Crocker SJ, Jackson-Lewis V, Jordan-Sciutto KL, Hayley S, Mount MP, et al. Cyclin-dependent kinase 5 is a mediator of dopaminergic neuron loss in a mouse model of Parkinson's disease. Proc Natl Acad Sci U S A. 2003;100(23):13650–5. doi:10.1073/pnas.2232515100.
142. Zhai D, Li S, Zhao Y, Lin Z. SLC6A3 is a risk factor for Parkinson's disease: a meta-analysis of sixteen years' studies. Neurosci Lett. 2014;564:99–104. doi:10.1016/j.neulet.2013.10.060.
143. Jacobs FM, van der Linden AJ, Wang Y, von Oerthel L, Sul HS, Burbach JP, et al. Identification of Dlk1, Ptpru and Klhl1 as novel Nurr1 target genes in meso-diencephalic dopamine neurons. Development. 2009;136(14):2363–73. doi:10.1242/dev.037556.
144. Okabe T, Takayanagi R, Imasaki K, Haji M, Nawata H, Watanabe T. cDNA cloning of a NGFI-B/nur77-related transcription factor from an apoptotic human T cell line. J Immunol. 1995;154(8):3871–9.
145. Xu PY, Liang R, Jankovic J, Hunter C, Zeng YX, Ashizawa T, et al. Association of homozygous 7048G7049 variant in the intron six of Nurr1 gene with Parkinson's disease. Neurology. 2002;58(6):881–4.
146. Carmine A, Buervenich S, Galter D, Jonsson EG, Sedvall GC, Farde L, et al. NURR1 promoter polymorphisms: Parkinson's disease, schizophrenia, and personality traits. Am J Med Genet B Neuropsychiatr Genet. 2003;120B(1): 51–7. doi:10.1002/ajmg.b.20033.
147. Le WD, Xu P, Jankovic J, Jiang H, Appel SH, Smith RG, et al. Mutations in NR4A2 associated with familial Parkinson disease. Nat Genet. 2003; 33(1):85–9. doi:10.1038/ng1066.
148. Tan EK, Chung H, Zhao Y, Shen H, Chandran VR, Tan C, et al. Genetic analysis of Nurr1 haplotypes in Parkinson's disease. Neurosci Lett. 2003;347(3):139–42.
149. Zheng K, Heydari B, Simon DK. A common NURR1 polymorphism associated with Parkinson disease and diffuse Lewy body disease. Arch Neurol. 2003;60(5):722–5. doi:10.1001/archneur.60.5.722.
150. Ibanez P, Lohmann E, Pollak P, Durif F, Tranchant C, Agid Y, et al. Absence of NR4A2 exon 1 mutations in 108 families with autosomal dominant Parkinson disease. Neurology. 2004;62(11):2133–4.
151. Levecque C, Destee A, Mouroux V, Amouyel P, Chartier-Harlin MC. Assessment of Nurr1 nucleotide variations in familial Parkinson's disease. Neurosci Lett. 2004;366(2):135–8. doi:10.1016/j.neulet.2004.05.028.
152. Nichols WC, Uniacke SK, Pankratz N, Reed T, Simon DK, Halter C, et al. Evaluation of the role of Nurr1 in a large sample of familial Parkinson's disease. Mov Disord. 2004;19(6):649–55. doi:10.1002/mds.20097.
153. Tan EK, Chung H, Chandran VR, Tan C, Shen H, Yew K, et al. Nurr1 mutational screen in Parkinson's disease. Mov Disord. 2004;19(12):1503–5. doi:10.1002/mds.20246.
154. Chu Y, Le W, Kompoliti K, Jankovic J, Mufson EJ, Kordower JH. Nurr1 in Parkinson's disease and related disorders. J Comp Neurol. 2006;494(3):495–514. doi:10.1002/cne.20828.
155. Grimes DA, Han F, Panisset M, Racacho L, Xiao F, Zou R, et al. Translated mutation in the Nurr1 gene as a cause for Parkinson's disease. Mov Disord. 2006;21(7):906–9. doi:10.1002/mds.20820.
156. Healy DG, Abou-Sleiman PM, Ahmadi KR, Gandhi S, Muqit MM, Bhatia KP, et al. NR4A2 genetic variation in sporadic Parkinson's disease: a genewide approach. Mov Disord. 2006;21(11):1960–3. doi:10.1002/mds.21018.
157. Chen CM, Chen IC, Chang KH, Chen YC, Lyu RK, Liu YT, et al. Nuclear receptor NR4A2 IVS6 + 18insG and brain derived neurotrophic factor (BDNF) V66M polymorphisms and risk of Taiwanese Parkinson's disease. Am J Med Genet B Neuropsychiatr Genet. 2007;144B(4):458–62. doi:10.1002/ajmg.b.30476.
158. Le W, Pan T, Huang M, Xu P, Xie W, Zhu W, et al. Decreased NURR1 gene expression in patients with Parkinson's disease. J Neurol Sci. 2008;273(1–2): 29–33. doi:10.1016/j.jns.2008.06.007.
159. Wu Y, Peng R, Chen W, Zhang J, Li T, Wang Y, et al. Association of the polymorphisms in NURR1 gene with Parkinson's disease. Zhonghua Yi Xue Yi Chuan Xue Za Zhi. 2008;25(6):693–6.
160. Sleiman PM, Healy DG, Muqit MM, Yang YX, Van Der Brug M, Holton JL, et al. Characterisation of a novel NR4A2 mutation in Parkinson's disease brain. Neurosci Lett. 2009;457(2):75–9. doi:10.1016/j.neulet.2009.03.021.

161. Lin X, Parisiadou L, Sgobio C, Liu G, Yu J, Sun L, et al. Conditional expression of Parkinson's disease-related mutant alpha-synuclein in the midbrain dopaminergic neurons causes progressive neurodegeneration and degradation of transcription factor nuclear receptor related 1. J Neurosci. 2012;32(27):9248–64. doi:10.1523/jneurosci.1731-12.2012.
162. Liu H, Wei L, Tao Q, Deng H, Ming M, Xu P, et al. Decreased NURR1 and PITX3 gene expression in Chinese patients with Parkinson's disease. Eur J Neurol. 2012;19(6):870–5. doi:10.1111/j.1468-1331.2011.03644.x.
163. Liu H, Tao Q, Deng H, Ming M, Ding Y, Xu P, et al. Genetic analysis of NR4A2 gene in a large population of Han Chinese patients with Parkinson's disease. Eur J Neurol. 2013;20(3):584–7. doi:10.1111/j.1468-1331.2012.03824.x.
164. Martin WE. Tyrosine hydroxylase deficiency. A unifying concept of Parkinsonism. Lancet. 1971;1(7708):1050–1.
165. Haavik J, Toska K. Tyrosine hydroxylase and Parkinson's disease. Mol Neurobiol. 1998;16(3):285–309. doi:10.1007/bf02741387.
166. Tabrez S, Jabir NR, Shakil S, Greig NH, Alam Q, Abuzenadah AM, et al. A synopsis on the role of tyrosine hydroxylase in Parkinson's disease. CNS Neurol Disord Drug Targets. 2012;11(4):395–409.
167. Zhu Y, Zhang J, Zeng Y. Overview of tyrosine hydroxylase in Parkinson's disease. CNS Neurol Disord Drug Targets. 2012;11(4):350–8.
168. Chandrasekaran S, Bonchev D. A network view on Parkinson's disease. Comput Struct Biotechnol J. 2013;7, e201304004. doi:10.5936/csbj.201304004.
169. Lin L, Isacson O. Axonal growth regulation of fetal and embryonic stem cell-derived dopaminergic neurons by Netrin-1 and Slits. Stem Cells. 2006;24(11):2504–13. doi:10.1634/stemcells.2006-0119.
170. Chatoo W, Abdouh M, David J, Champagne MP, Ferreira J, Rodier F, et al. The polycomb group gene Bmi1 regulates antioxidant defenses in neurons by repressing p53 pro-oxidant activity. J Neurosci. 2009;29(2):529–42. doi:10.1523/jneurosci.5303-08.2009.
171. Thomas B, Beal MF. Parkinson's disease. Hum Mol Genet. 2007;16 Spec No. 2:R183–94. doi:10.1093/hmg/ddm159.
172. Johnson MT, Yang HS, Magnuson T, Patel MS. Targeted disruption of the murine dihydrolipoamide dehydrogenase gene (Dld) results in perigastrulation lethality. Proc Natl Acad Sci U S A. 1997;94(26):14512–7.
173. Klivenyi P, Starkov AA, Calingasan NY, Gardian G, Browne SE, Yang L, et al. Mice deficient in dihydrolipoamide dehydrogenase show increased vulnerability to MPTP, malonate and 3-nitropropionic acid neurotoxicity. J Neurochem. 2004;88(6):1352–60.
174. Jenner P, Marsden CD. The actions of 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine in animals as a model of Parkinson's disease. J Neural Transm Suppl. 1986;20:11–39.
175. Gibson GE, Park LC, Sheu KF, Blass JP, Calingasan NY. The alpha-ketoglutarate dehydrogenase complex in neurodegeneration. Neurochem Int. 2000;36(2):97–112.
176. Gibson GE, Kingsbury AE, Xu H, Lindsay JG, Daniel S, Foster OJ, et al. Deficits in a tricarboxylic acid cycle enzyme in brains from patients with Parkinson's disease. Neurochem Int. 2003;43(2):129–35.
177. Mizuno Y, Suzuki K, Ohta S. Postmortem changes in mitochondrial respiratory enzymes in brain and a preliminary observation in Parkinson's disease. J Neurol Sci. 1990;96(1):49–57.
178. Papapetropoulos S, Ffrench-Mullen J, McCorquodale D, Qin Y, Pablo J, Mash DC. Multiregional gene expression profiling identifies MRPS6 as a possible candidate gene for Parkinson's disease. Gene Expr. 2006;13(3):205–15.
179. Papapetropoulos S, Ffrench-Mullen J, Mash DC. Gene expression profiling of Parkinson's Disease. Google Patents. 2012.
180. Sgado P, Ferretti E, Grbec D, Bozzi Y, Simon HH. The atypical homeoprotein Pbx1a participates in the axonal pathfinding of mesencephalic dopaminergic neurons. Neural Dev. 2012;7:24. doi:10.1186/1749-8104-7-24.
181. Plante-Bordeneuve V, Taussig D, Thomas F, Said G, Wood NW, Marsden CD, et al. Evaluation of four candidate genes encoding proteins of the dopamine pathway in familial and sporadic Parkinson's disease: evidence for association of a DRD2 allele. Neurology. 1997;48(6):1589–93.
182. Pastor P, Munoz E, Obach V, Marti MJ, Blesa R, Oliva R, et al. Dopamine receptor D2 intronic polymorphism in patients with Parkinson's disease. Neurosci Lett. 1999;273(3):151–4.
183. Costa-Mallen P, Costa LG, Smith-Weller T, Franklin GM, Swanson PD, Checkoway H. Genetic polymorphism of dopamine D2 receptors in Parkinson's disease and interactions with cigarette smoking and MAO-B intron 13 polymorphism. J Neurol Neurosurg Psychiatry. 2000;69(4):535–7.
184. Grevle L, Guzey C, Hadidi H, Brennersted R, Idle JR, Aasly J. Allelic association between the DRD2 TaqI A polymorphism and Parkinson's disease. Mov Disord. 2000;15(6):1070–4.
185. Oliveri RL, Annesi G, Zappia M, Civitelli D, De Marco EV, Pasqua AA, et al. The dopamine D2 receptor gene is a susceptibility locus for Parkinson's disease. Mov Disord. 2000;15(1):127–31.
186. Kelada SN, Costa-Mallen P, Costa LG, Smith-Weller T, Franklin GM, Swanson PD, et al. Gender difference in the interaction of smoking and monoamine oxidase B intron 13 genotype in Parkinson's disease. Neurotoxicology. 2002;23(4–5):515–9.
187. Tan EK, Tan Y, Chai A, Tan C, Shen H, Lum SY, et al. Dopamine D2 receptor TaqIA and TaqIB polymorphisms in Parkinson's disease. Mov Disord. 2003;18(5):593–5. doi:10.1002/mds.10406.
188. Singh M, Khan AJ, Shah PP, Shukla R, Khanna VK, Parmar D. Polymorphism in environment responsive genes and association with Parkinson disease. Mol Cell Biochem. 2008;312(1–2):131–8. doi:10.1007/s11010-008-9728-2.
189. Lee JY, Lee EK, Park SS, Lim JY, Kim HJ, Kim JS, et al. Association of DRD3 and GRIN2B with impulse control and related behaviors in Parkinson's disease. Mov Disord. 2009;24(12):1803–10. doi:10.1002/mds.22678.
190. Kiyohara C, Miyake Y, Koyanagi M, Fujimoto T, Shirasawa S, Tanaka K, et al. Genetic polymorphisms involved in dopaminergic neurotransmission and risk for Parkinson's disease in a Japanese population. BMC Neurol. 2011;11:89. doi:10.1186/1471-2377-11-89.
191. McGuire V, Van Den Eeden SK, Tanner CM, Kamel F, Umbach DM, Marder K, et al. Association of DRD2 and DRD3 polymorphisms with Parkinson's disease in a multiethnic consortium. J Neurol Sci. 2011;307(1–2):22–9. doi:10.1016/j.jns.2011.05.031.
192. Lee JY, Cho J, Lee EK, Park SS, Jeon BS. Differential genetic susceptibility in diphasic and peak-dose dyskinesias in Parkinson's disease. Mov Disord. 2011;26(1):73–9. doi:10.1002/mds.23400.
193. Kumudini N, Umai A, Devi YP, Naushad SM, Mridula R, Borgohain R, et al. Impact of COMT H108L, MAOB int 13 A > G and DRD2 haplotype on the susceptibility to Parkinson's disease in South Indian subjects. Indian J Biochem Biophys. 2013;50(5):436–41.
194. Dai D, Wang Y, Wang L, Li J, Ma Q, Tao J, et al. Polymorphisms of and genes and Parkinson's disease: A meta-analysis. Biomed Rep. 2014;2(2):275–81. doi:10.3892/br.2014.220.
195. Sgado P, Alberi L, Gherbassi D, Galasso SL, Ramakers GM, Alavian KN, et al. Slow progressive degeneration of nigral dopaminergic neurons in postnatal Engrailed mutant mice. Proc Natl Acad Sci U S A. 2006;103(41):15242–7. doi:10.1073/pnas.0602116103.
196. Le Pen G, Sonnier L, Hartmann A, Bizot JC, Trovero F, Krebs MO, et al. Progressive loss of dopaminergic neurons in the ventral midbrain of adult mice heterozygote for Engrailed1: a new genetic model for Parkinson's disease? Parkinsonism Relat Disord. 2008;14 Suppl 2:S107–11. doi:10.1016/j.parkreldis.2008.04.007.
197. Haubenberger D, Reinthaler E, Mueller JC, Pirker W, Katzenschlager R, Froehlich R, et al. Association of transcription factor polymorphisms PITX3 and EN1 with Parkinson's disease. Neurobiol Aging. 2011;32(2):302–7. doi:10.1016/j.neurobiolaging.2009.02.015.
198. Zoni S, Bonetti G, Lucchini R. Olfactory functions at the intersection between environmental exposure to manganese and Parkinsonism. J Trace Elem Med Biol. 2012;26(2–3):179–82. doi:10.1016/j.jtemb.2012.04.023.
199. Rosenbaum JN, Duggan A, Garcia-Anoveros J. Insm1 promotes the transition of olfactory progenitors from apical and proliferative to basal, terminally dividing and neuronogenic. Neural Dev. 2011;6:6. doi:10.1186/1749-8104-6-6.
200. Duggan A, Madathany T, de Castro SC, Gerrelli D, Guddati K, Garcia-Anoveros J. Transient expression of the conserved zinc finger gene INSM1 in progenitors and nascent neurons throughout embryonic and adult neurogenesis. J Comp Neurol. 2008;507(4):1497–520. doi:10.1002/cne.21629.
201. Westermann B, Wattendorf E, Schwerdtfeger U, Husner A, Fuhr P, Gratzl O, et al. Functional imaging of the cerebral olfactory system in patients with Parkinson's disease. J Neurol Neurosurg Psychiatry. 2008;79(1):19–24. doi:10.1136/jnnp.2006.113860.
202. Haehner A, Hummel T, Reichmann H. Olfactory dysfunction as a diagnostic marker for Parkinson's disease. Expert Rev Neurother. 2009;9(12):1773–9. doi:10.1586/ern.09.115.
203. Wattendorf E, Welge-Lussen A, Fiedler K, Bilecen D, Wolfensberger M, Fuhr P, et al. Olfactory impairment predicts brain atrophy in Parkinson's disease. J Neurosci. 2009;29(49):15410–3. doi:10.1523/jneurosci.1909-09.2009.

204. Teunissen CE, Veerhuis R, De Vente J, Verhey FR, Vreeling F, van Boxtel MP, et al. Brain-specific fatty acid-binding protein is elevated in serum of patients with dementia-related diseases. Eur J Neurol. 2011;18(6):865–71. doi:10.1111/j.1468-1331.2010.03273.x.
205. Watanabe A, Toyota T, Owada Y, Hayashi T, Iwayama Y, Matsumata M, et al. Fabp7 maps to a quantitative trait locus for a schizophrenia endophenotype. PLoS Biol. 2007;5(11), e297. doi:10.1371/journal.pbio.0050297.
206. Grauer SM, Hodgson R, Hyde LA. MitoPark mice, an animal model of Parkinson's disease, show enhanced prepulse inhibition of acoustic startle and no loss of gating in response to the adenosine A(2A) antagonist SCH 412348. Psychopharmacology. 2014;231(7):1325–37. doi: 10.1007/s00213-013-3320-5.
207. Zoetmulder M, Biernat HB, Nikolic M, Korbo L, Friberg L, Jennum PJ. Prepulse Inhibition is Associated with Attention, Processing Speed, and < sup > 123</sup > I-FP-CIT SPECT in Parkinson's Disease. J Parkinsons Dis. 2014;4(1):77–87. doi:10.3233/jpd-130307.
208. Fung HC, Scholz S, Matarin M, Simon-Sanchez J, Hernandez D, Britton A, et al. Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. Lancet Neurol. 2006;5(11):911–6. doi:10.1016/s1474-4422(06)70578-6.
209. Kitamura N, Hashimoto T, Nishino N, Tanaka C. Inositol 1,4,5-trisphosphate binding sites in the brain: regional distribution, characterization, and alterations in brains of patients with Parkinson's disease. J Mol Neurosci. 1989;1(3):181–7.
210. Ding J, Guzman JN, Tkatch T, Chen S, Goldberg JA, Ebert PJ, et al. RGS4-dependent attenuation of M4 autoreceptor function in striatal cholinergic interneurons following dopamine depletion. Nat Neurosci. 2006;9(6):832–42. doi:10.1038/nn1700.
211. Lerner TN, Kreitzer AC. RGS4 is required for dopaminergic control of striatal LTD and susceptibility to parkinsonian motor deficits. Neuron. 2012;73(2): 347–59. doi:10.1016/j.neuron.2011.11.015.
212. Ko WK, Martin-Negrier ML, Bezard E, Crossman AR, Ravenscroft P. RGS4 is involved in the generation of abnormal involuntary movements in the unilateral 6-OHDA-lesioned rat model of Parkinson's disease. Neurobiol Dis. 2014;70:138–48. doi:10.1016/j.nbd.2014.06.013.
213. Maraganore DM, Wilkes K, Lesnick TG, Strain KJ, de Andrade M, Rocca WA, et al. A limited role for DJ1 in Parkinson disease susceptibility. Neurology. 2004;63(3):550–3.