Neuroendocrinology Letters
ISSN: 0172-780X; ISSN-L: 0172-780X; Electronic/Online ISSN: 2354-4716
*Web of Knowledge / Web of Science:* Neuroendocrinol Lett
*Pub Med / Medline:* Neuro Endocrinol Lett

ORIGINAL ARTICLE

# Identifying distinct candidate genes for early Parkinson's disease by analysis of gene expression in whole blood

**Ai-Guo Sun, Jing Wang, Yao-Zhong Shan, Wen-Jiao Yu, Xin Li, Chao-Hua Cong, Xin Wang**

Department of Neurology, The First Affiliated Hospital of Harbin Medical University, Harbin 150001, PR China

*Correspondence to:* Ai-Guo Sun
Department of Neurology,
The First Affiliated Hospital of Harbin Medical University
No. 23, Youzheng Street, Nangang District Harbin,
Heilongjiang Province, China, 150001.
tel: +86-0451-85555668; fax: +86-0451-53605867; e-mail: Sjnkwbf1975@163.com

*Submitted: 2014-04-27 Accepted: 2014-06-21 Published online: 2014-09-28*

*Key words:* **Parkinson's Disease; microarray; PCA; functional analysis; pathway; distinct candidate**

Neuroendocrinol Lett 2014; **35**(5):398–404 **PMID:** 25275262 NEL350514A09 

## Abstract

**OBJECTIVE:** Parkinson disease (PD) is a degenerative disorder of the central nervous system, and in the majority of cases, the causes of PD are unknown. Coupled with impressive advances in statistical tools for analyzing large, complex data sets, well-designed microarray experiments are poised to make a big impact in the field of diseases. So we set the study to identify distinct PD-associated candidates.

**METHODS:** Candidate genes, with statistical significant changes of expression in PD patients' samples, were extracted from a transcriptome-wide microarray data in 105 individuals, which were downloaded from GEO, NCBI, by using statistical methods; Selected findings were confirmed by principal component analysis (PCA) and functional and pathway enrichment analysis were used to further study about the distinct candidates.

**RESULTS:** A total of 10 distinctly differentially expressed genes were identified in PD patitents' samples. After PCA confirmation, we specifically pointed out 4 genes (PRKAG2, DLG1, DDX3Y, RPS4Y) as the high confidence distinct candidates in PD. Network and functional categories showed that they were most related to translational elongation(GO:0006414) and participated in mTOR signaling pathway(hsa04150).

**CONCLUSION:** Among 10 distinct genes which are identified in PD patients' samples, DLG1, XIST, DDX3Y and RPS4Y1 genes can classify samples into different group clearly, and they are regarded as high confidence distinct gene biomarkers of PD. Our results provide a systematic view of the functional alterations of PD that may help to elucidate the mechanisms of PD and lead to improved treatments for PD patients.

**Abbreviations:**

| | |
|---|---|
| PD | - Parkinson disease |
| PCA | - Principal component analysis |
| ND | - Neurodegenerative diseases |
| GEO | - Gene Expression Omnibus |
| NCBI | - National Center of Biotechnology Information |
| mTOR | - Mammalian Target Of Rapamycin |
| PRKAG2 | - 5′-AMP-activated protein kinase subunit gamma-2 |
| RPS4Y | - 40S ribosomal protein S4, Y isoform 1 |
| DLG1 | - Disks large homolog 1 |
| XIST | - X inactive specific transcript |
| DDX3Y | - DEAD (Asp-Glu-Ala-Asp) box helicase 3, Y-linked |
| RMA | - Robust multiarray average |
| FC | - Fold change |
| PPIs | - Protein-protein interactions |
| EASE | - Expression Analysis Systematic Explorer software |
| HC | - Healthy control |

## INTRODUCTION

Parkinson's disease is a gradually progressive, degenerative neurologic disorder which typically impairs the patient's motor skills, speech, writing, as well as some other functions (Jankovic 2008; Aarsland *et al.* 2003). Expression profiling of mRNA has been also used to study about various types of diseases. A range of gene signatures which have important roles as bio-markers or target gene in diseases, have been identified by the application of DNA chips (Huang *et al.* 2011). The development of biomarkers for PD would have tremendous utility. It may prove to be useful in identifying at risk individuals, or in early diagnosis and in identifying subgroups of PD. The remarkable progress made by molecular biology and molecular genetics during the past decades, and the advent of the novel tools of genomics and proteomics, are expected to reveal differential expression profiles of thousands of genes and proteins involved in Parkinson's disease. Of particular interest is the application of microarrays in drug discovery and design to potential candidate targets for medicine intervention.

Here, we identified statistical significant changes of expression in PD patients' samples, were extracted from a transcriptome-wide microarray data in 105 individuals, which were downloaded from GEO, NCBI, by using statistical methods; Selected findings were confirmed by principal component analysis (PCA) and functional and pathway enrichment analysis were used to further study about the distinct candidates. The elucidation of important gene expression patterns during disease will make possible identification of genetic susceptibility markers, biomarkers of disease progression, and new therapeutic targets.

## MATERIALS AND METHODS

### *Data source*

We downloaded the gene expression profiles of whole blood from 50 patients with PD, 33 with neurodegenerative diseases (ND) other than PD, and 23 healthy control samples (Scherzer *et al.* 2007) from GEO (Gene Expression Omnibus) with the accession number GSE6613, a set of U133A chips (together representing 22,283 probe sets). We regarded healthy and other neurodegenerative diseases samples as controls of PD ones. All of the studies were approved by the Human Ethics Committee of data providers, Technical University of Denmark.

### *Data processing and signifigance analysis*

DNA microarray expression profile data of all the PD and control samples were normalized simultaneously using robust multiarray average (RMA) method (Best *et al.* 2005), and to identify genes differently expressed in relation with PD, filtered data were analyzed with the test method in Limma package (Smyth *et al.* 2003; Smyth 2005), implemented in R language, Bioconductor project. Meanwhile, fold change between groups were also calculated, because each sample may show intrinsic individual variability, the threshold for determining the fold change (FC) was set at 2. Differentially expressed genes were those defined with the cutoff of $p$-value <0.05 and $|\log FC|>1$. The $p$-value or FDR less than 0.05 was considered as statistically significant in our study.

### *Hierarchical clustering analysis of selected genes*

The difference expression genes between patients and controls were used to generate hierarchical clustering image by CLUSTER3.0 (Yeung & Ruzzo 2001), using Pearson correlation (uncentered), complete linkage clustering (Jain & Dubes 1988), with normalized data, and visualized the hierarchical clustering heat-map with TreeView (Eisen *et al.* 1998).

### *Selection of distinct gene biomarkers*

We compared differentially expressed genes, selected from PD VS. healthy and ND VS. healthy, found the common and specific part of each group, Student's t-test (Sawilosky 2005) was used to compare common genes of groups. To construct a predictive model of the changes observed, a projection by principal component analysis (PCA) (Abdi & Williams 2010) was carried out.

### *Network construction of distinct gene biomarkers*

Protein-protein interactions (PPIs) are crucial for all biological processes, and provide a valuable framework for a better understanding of the functional organization of the proteome (Stelzl *et al.* 2005). To detect interacting pairs of our selected biomarkers, we constructed interaction network by using String database (Snel *et al.* 2005; von Mering *et al.* 2003).

### *Functional and KEGG pathway analysis of the network genes*

To further analysis of functions of network genes, the approach consisted in assigning genes in the interaction network to biological and participated KEGG

pathways by using the hierarchical database of the Gene Ontology (GO) consortium (The Gene Ontology Consortium 2008; Diehl *et al.* 2007) and EASE (Ford *et al.* 2006) (Expression Analysis Systematic Explorer software) with a cutoff of FDR lower than 0.05.

## RESULTS

As shown in Figure 1B, a multi-step flow was adopted to identify the specific genes expressed in the PD patients samples. Firstly, the significantly aberrant expression profiles of genes were obtained from PD VS. healthy and ND VS healthy groups, by statistical tests with R language packages; Secondly, PCA was carried out to confirm the most specific bio-marker genes, and constructed regulation network with help of databases; Thirdly, found enriched functional GO terms and related pathways.

*Signifigance analysis and clustering*

A total of 50 patients with PD, 33 with neurodegenerative diseases(ND) other than PD, and 23 healthy control(HC) samples were enrolled in our study, shown in Figure 1A, and we divided the samples into two groups: PD VS. HC and ND VS. HC. After between-array normalization and filtering carried out by RMA algorithm and using statistical test in Limma, 11 and 26 genes were found to be significantly differentially expressed in PD VS. HC and ND VS. HC group, respectively, with a cutoff of $p$-value <0.05. In addition, the heatmaps of aberrantly different gene levels in two groups were demonstrated in Figure 2.

*Selection of distinct gene biomarkers*

We compared the 11 and 26 selected genes which were significantly differentially expressed in PD VS. healthy and ND VS healthy group, only one – gene PCDH7, was the common part of two groups, as shown in Figure 3A, expression of gene PCDH7 were significantly expressed in both of groups: PD VS. HC ($p$=0.0006) and ND VS. HC ($p$=0.013), and it was over-expressed in PD and ND patients (Figure 3B). Because of the similar expression pattern of gene PCDH in two groups, it can't be the specific part of PD VS. HC, so we ticked it out. That is to say, apart from the same part with ND, there were 10 specific expression genes in PD patients samples, we regarded them as the candidate biomarker genes for the further confirmation.

*Confirmation of candidate biomarker genes*

Principal component analysis of the the above 10 selected genes in PD patients samples, which were identified as differentially expressed in PD VS. HC by SAM, confirmed that they could clearly separate PD and healthy individuals, as shown in Figure 4. The first (Figure 4A, axis 1) and second (Figure 4B, axis 2) factors can distinguish PD samples from healthy ones clearly. Genes with a high correlation coefficient were considered to be important for discriminating the two groups (Figure 3B). Genes like DLG1, XIST, DDX3Y

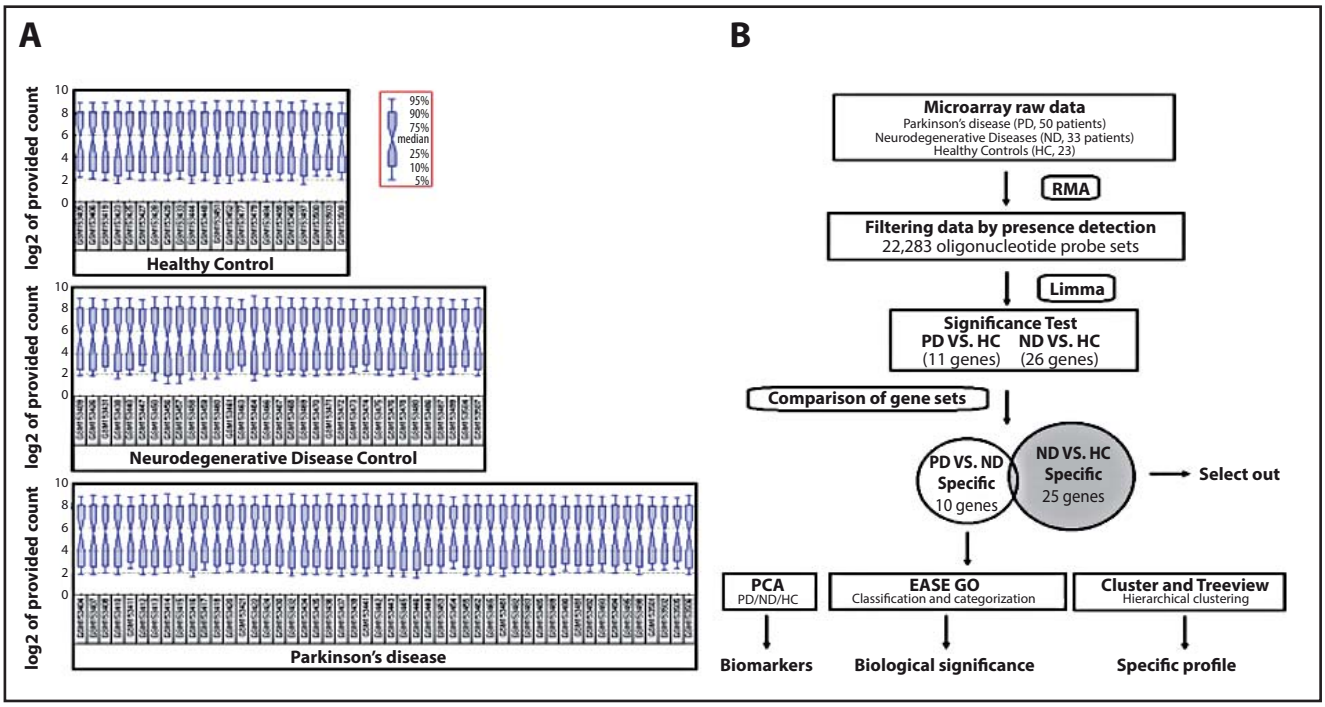


**Fig. 1.** Samples and multi-step flowchart. A). Samples in each group: Healthy Control(HC), neurodegenerative disease(ND) control and parkinson's disease(PD). B). Flow chart. Step 1, mRNA microarray data processing and tests for identifying significantly differentially expressed genes in group PD VS. HC and ND VS. HC; Step 2, compare differentially expressed genes selected from group group PD VS. HC and ND VS. HC, get the common and specific part of each gene set; Step 3, specific genes in PD VS.HC are analyzed by PCA, , investigated by clustering and classified by GO functional categories.

and RPS4Y1, which were located in the upper side of the coordinate, characterized the group of PD patient samples.

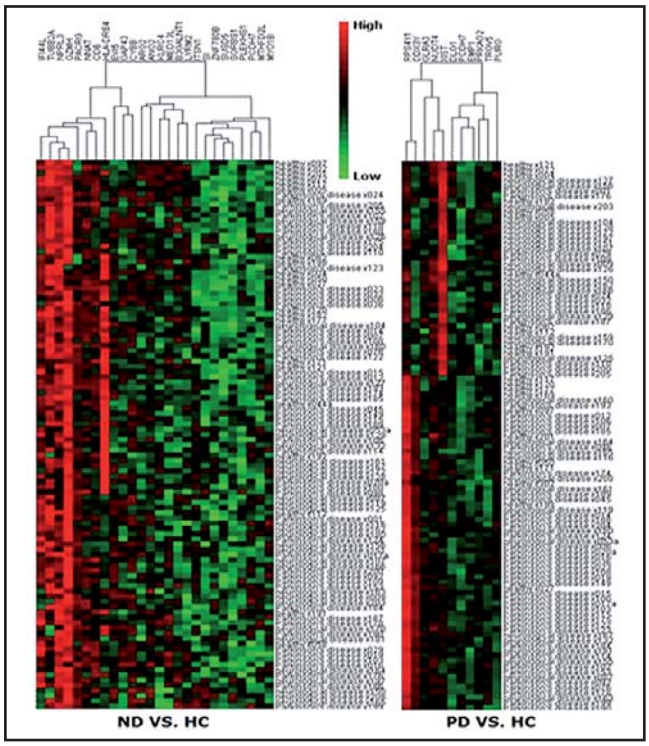


**Fig. 2.** Heatmap of selected genes expression in each sample and group, ND VS. HC(left), PD VS. HC (right). The color means expression levels: green means low expression and red means high expression.

*Network construction of distinct gene biomarkers*

DLG1, XIST, DDX3Y and RPS4Y1 genes are not only differentially expressed in PD patients, but also confirmed by PCA, so we regarded these genes as the distinct gene biomarkers of PD patients. In order to search related interactors of them, we constructed interaction

**Tab. 1.** Enriched GO term list.

| GO Term | Gene count | FDR |
|---|---|---|
| GO:0006414~translational elongation | 29 | 1.14E-35 |
| GO:0003735~structural constituent of ribosome | 29 | 6.39E-29 |
| GO:0006412~translation | 31 | 4.78E-23 |
| GO:0046320~regulation of fatty acid oxidation | 13 | 1.75E-16 |
| GO:0019217~regulation of fatty acid metabolic process | 14 | 7.86E-15 |
| GO:0010565~regulation of cellular ketone metabolic process | 14 | 7.00E-14 |
| GO:0005198~structural molecule activity | 30 | 5.88E-14 |
| GO:0019216~regulation of lipid metabolic process | 14 | 7.35E-10 |
| GO:0003723~RNA binding | 24 | 3.18E-07 |
| GO:0019899~enzyme binding | 18 | 8.38E-05 |
| GO:0019901~protein kinase binding | 10 | 8.31E-04 |
| GO:0019900~kinase binding | 10 | 4.29E-03 |

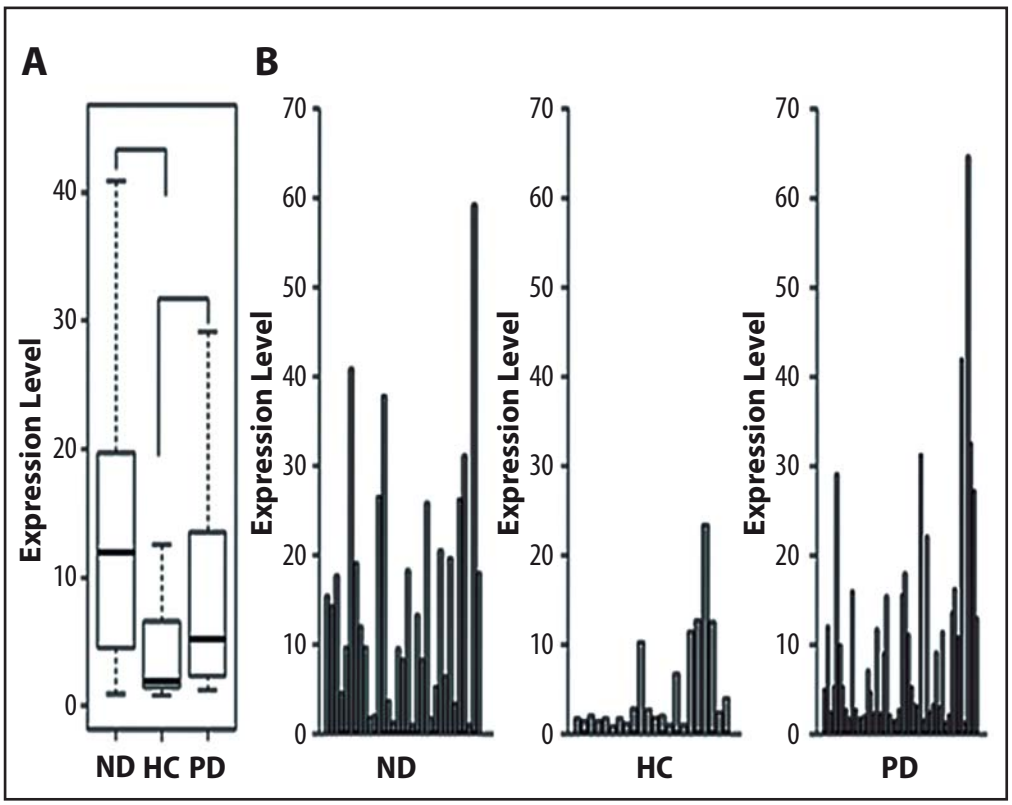


**Fig. 3.** Expression of gene PCDH7. A). Box plot of PCDH7 expression in each group. B). Bar plot of PCDH7 expression in each sample and group.

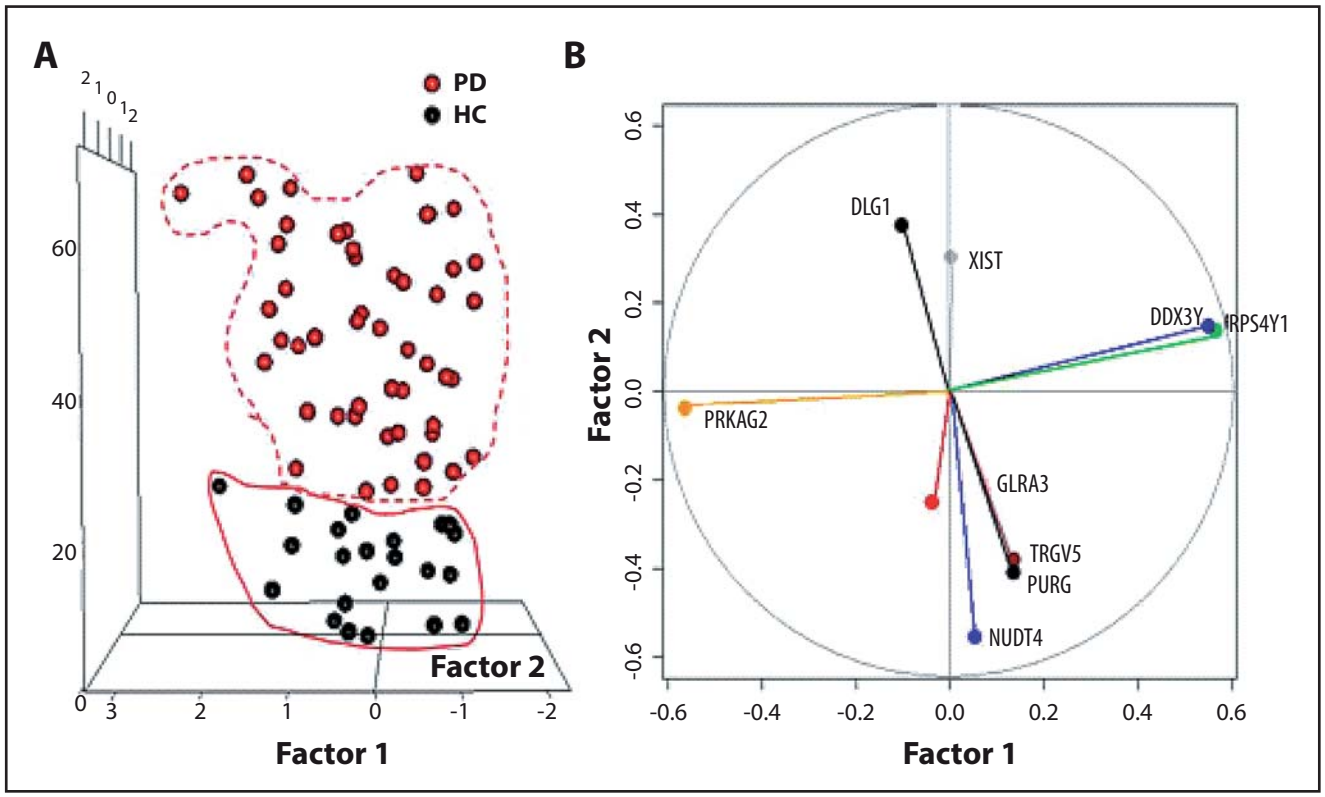


**Fig. 4.** Principal component analysis of the 10 genes that are differentially regulated in PD samples detected by Limma. A). Projection of the individuals shows that PD patient samples can be discriminated from the healthy ones; B). Projection of the genes. Genes located located in the upper side of the coordinate, characterized the group of PD patient samples.

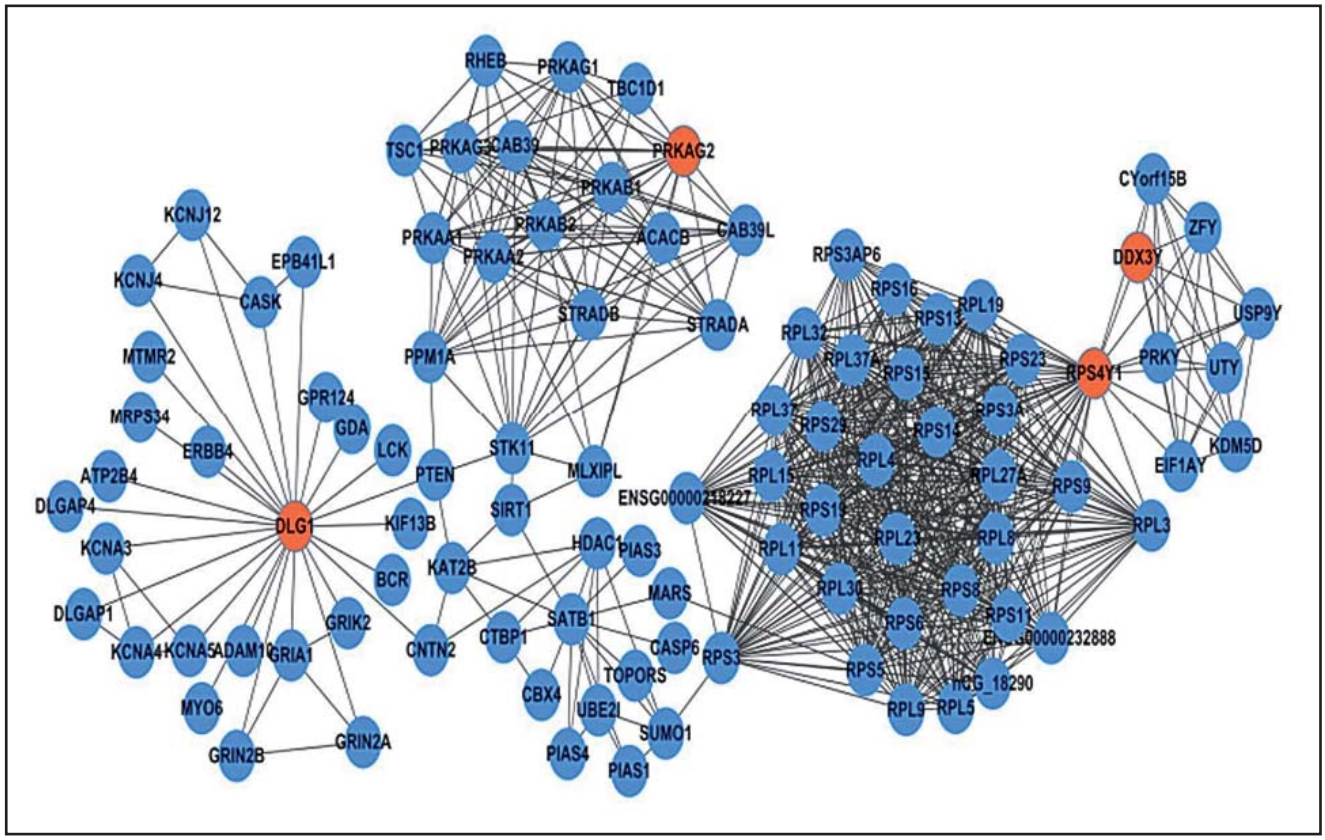


**Fig. 5.** Interaction network of distinct gene biomarkers. The orange and blue nodes represent our selected genes and the predictive interactors, respectively.

**Tab. 2.** Significant KEGG pathways list.

| Term | Count | FDR | Genes |
|---|---|---|---|
| hsa04150:mTOR signaling pathway | 9 | 3.05E-04 | CAB39L, TSC1, STK11, STRADA, RHEB,PRKAA1, CAB39, PRKAA2, RPS6 |
| hsa04920:Adipocytokine signaling pathway | 9 | 0.00226 | PRKAG3, STK11, PRKAG1, PRKAB2, PRKAG2, PRKAB1, PRKAA1, ACACB, PRKAA2 |
| hsa04910:Insulin signaling pathway | 11 | 0.0094 | PRKAG3, TSC1, PRKAG1, PRKAB2, PRKAG2, PRKAB1, RHEB, PRKAA1, ACACB, PRKAA2, RPS6 |

network with String database. As shown in Figure 5, we finally got a network consist of distinct gene biomarkers and their predictive interactors.

*Functional and KEGG pathway analysis of the network genes*

With the help of GO annotations and EASE software, we searched the significance related GO term and KEGG pathways based on the enrichment algorithm. We finally obtained 12 significantly enriched GO function terms, just as listed in Table 1, GO:0006414: translational elongation with the lowest FDR value, was the most related function of the network genes. Meanwhile, 3 significant KEGG pathways were selected and shown in Table 2, and the network genes were most significantly enriched in the mTOR signaling pathway (hsa04150).

## DISCUSSION

Identification of biomarkers for PD is an important step towards improving current diagnostic criteria, identifying at risk individuals and disease subgroups. This is important, since clinical criteria are at best 90% accurate, and atypical parkinsonian disorders, such as multiple system atrophy and progressive supranuclear palsy, are generally unresponsive to pharmacotherapy and surgical treatment. Additionally, biomarkers could provide insights into disease mechanisms, which in turn, could be used to identify aberrant biochemical pathways and therapeutic targets and to develop efficacious medications. So the development of biomarkers for PD has great potential significance for clinical.

In our study, we used a multiple-step approach to identify potential biomarkers for PD. Among 22238 probes, we have identified 11 genes with significant expression changes in PD patients' samples. We then compared the list of genes that selected from patients who suffered from ND other than PD, and found only one common gene. The left 10 genes are regarded as high confidence distinct gene biomarkers of PD after PCA confirmation, especially DLG1, XIST, DDX3Y and RPS4Y1 genes, and they can classify samples into different group clearly. In order to search related interactors of them, we constructed interaction network with String database, and genes in the network are most related to translational elongation(GO:0006414) and participated in mTOR signaling pathway(hsa04150). Previous studies showed that translational elongation has close relationship with PD, such as translation elongation factor 1A (Inamura *et al.* 2005; Gross & Kinzy 2005). What's more, the mammalian target of rapamycin (mTOR) pathway is an essential cellular signaling pathway involved in a number of important physiological functions, including cell growth, proliferation, metabolism, protein synthesis, and autophagy. Dysregulation of the mTOR pathway has been implicated in the pathophysiology of a number of neurological diseases, such as PD (Laplante & Sabatini 2012; Weber & Gutmann 2012; Cho 2011). So we conclude that our selected genes and their interactors enriched in such GO term are closely related to PD. Our results provide a systematic view of the functional alterations of PD that may help to elucidate the mechanisms of PD and lead to improved treatments for PD patients, but more work is needed.

## ACKNOWLEDGEMENT

All the data we used in the study are downloaded from public database, NCBI. This work was supported by Grants from the Heilongjiang Health Bureau (No.2010-023). We sincerely thank provider of microarray data, Technical University of Denmark, and all the authors who contributed to this article.

REFERENCES

1 Aarsland D, Andersen K, Larsen JP, Lolk A, Kragh-Sorensen P(2003). Prevalence and characteristics of dementia in Parkinson disease: an 8-year prospective study. Arch Neurol. **60**: 387–392.
2 Abdi H, Williams LJ (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. **2**: 433–459.
3 Best CJ, Gillespie JW, Yi Y, *et al.* (2005). Molecular alterations in primary prostate cancer after androgen ablation therapy. Clin. Cancer Res. **11**: 6823–6834.
4 Cho CH (2011). Frontier of epilepsy research-mTOR signaling pathway. Exp Mol Med. **43**: 231–274.
5 Diehl AD, Lee JA, Scheuermann RH, Blake JA (2007). Ontology development for biological systems: immunology. Bioinformatics. **23**(7): 913–915.
6 Eisen MB, Spellman PT, Brown PO, and Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA. **95**: 14863–14868.
7 Ford G, Xu Z, Gates A, Jiang J, Ford BD (2006). Expression Analysis Systematic Explorer (EASE) analysis reveals differential gene expression in permanent and transient focal stroke rat models. Brain Res. **1071**(1): 226–236.

8 Gross SR, Kinzy TG (2005). Translation elongation factor 1A is essential for regulation of the actin cytoskeleton and cell morphology. Nat Struct Mol Biol. **12**(9): 772–778.
9 Huang T, Wan S, Xu Z, Zheng Y, Feng KY, *et al.* (2011). Analysis and prediction of translation rate based on sequence and functional features of the mRNA. PLoS ONE **6**: e16036.
10 Inamura N, Nawa H, Takei NJ. Neurochem (2005).Enhancement of translation elongation in neurons by brain-derived neurotrophic factor: implications for mammalian target of rapamycin signaling. J Neurochem. **95**(5): 1438–1445.
11 Jain AK, and Dubes RC (1988). Algorithms for clustering data (Englewood Cliffs, N.J.: Prentice Hall).
12 Jankovic J (2008). Parkinson's disease: clinical features and diagnosis. J. Neurol. Neurosurg. Psychiatr. **79**(4): 368–376.
13 Laplante M, Sabatini DM (2012). mTOR Signaling in Growth Control and Disease.Cell. **149**(2): 274–293.
14 Sawilowsky S (2005). Misconceptions leading to choosing the t test over the Wilcoxon Mann-Whitney U test for shift in location parameter. Journal of Modern Applied Statistical Methods. **4**(2): 598–600.
15 Scherzer CR, Eklund AC, Morse LJ, Liao Z *et al.* (2007). Molecular markers of early Parkinson's disease based on gene expression in blood. Proc Natl Acad Sci USA. **104**(3): 955–960.
16 Smyth GK (2005). Limma: linear models for microarray data. In 'Bioinformatics and Computational Biology Solutions using R and Bioconductor'. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, 2005.
17 Smyth GK, Yang Y-H, Speed T P (2003). Statistical issues in microarray data analysis. Methods in Molecular Biology. **224**: 111–136.
18 Snel LJ, Hooper B, Krupp SD, Huynen MA, *et al.* (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Res. **33**(Database issue): D433–437.
19 Stelzl U, Worm U, Lalowski M, Haenig C, *et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell. **122**(6): 957–968.
20 The Gene Ontology Consortium (2008). The Gene Ontology project in 2008. Nucleic Acids Res. **36**(Database issue): D440–444.
21 von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003). STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. **31**(1): 258–261.
22 Weber JD,Gutmann DH (2012). Deconvoluting mTOR biology. Cell Cycle. **11**: 236–248.
23 Yeung KY, and Ruzzo WL (2001). Principal Component Analysis for clustering gene expression data. Bioinformatics. **17**: 763–774.