Database, Vol. 2011, Article ID bar007, doi:10.1093/database/bar007

Database tool

ParkDB: a Parkinson's disease gene expression database

Cristian Taccioli[1], Vincenza Maselli[1], Jesper Tegnér[2], David Gomez-Cabrero[2], Gioia Altobelli[3], Warren Emmett[4], Francesco Lescai[5], Stefano Gustincich[6] and Elia Stupka[1,7,*]

[1]UCL, Department of Cancer Biology, University College London, Gower Street, London, UK, [2]Department of Medicine, Unit of Computational Medicine, Centre of Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden, [3]Department for Endocrinology, William Harvey Research Institute, Queen Mary University of London, Charterhouse Square, London, UK, [4]CBM S.c.r.l., Basovizza, Trieste, Italy, [5]Division of Research Strategy, Faculty of Biomedical Services, University College London, Gower Street, London, UK, [6]Department of Neurobiology, International School for Advanced Studied (S.I.S.S.A.-I.S.A.S.), Area Science Park, Ss 14, Km 163.5, Trieste, Italy and [7]Blizard Institute of Cell and Molecular Science, Barts and The London School of Medicine and Dentistry, 4 Newark Street, Whitechapel, London, UK

***Corresponding author:** Tel.: +44 20 7679 6493, Fax: +44 20 7679 6817 Email: e.stupka@ucl.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

Submitted 14 November 2010; Revised 11 February 2011; Accepted 9 March 2011

Parkinson's disease (PD) is a common, adult-onset, neuro-degenerative disorder characterized by the degeneration of cardinal motor signs mainly due to the loss of dopaminergic neurons in the substantia nigra. To date, researchers still have limited understanding of the key molecular events that provoke neurodegeneration in this disease. Here, we present ParkDB, the first queryable database dedicated to gene expression in PD. ParkDB contains a complete set of re-analyzed, curated and annotated microarray datasets. This resource enables scientists to identify and compare expression signatures involved in PD and dopaminergic neuron differentiation under different biological conditions and across species.

Database URL: http://www2.cancer.ucl.ac.uk/Parkinson_Db2/

Introduction

Parkinson's disease (PD) is both sporadic and genetic (5–10%) (1). Although several mutated loci have been identified, they appear to be directly responsible in only a relatively small number of families. Evidence is accumulating, however, that the molecular pathways underlying these genomic regions may be common to more than one genetic form of Parkinsonism and may also play a role in the common sporadic disease (2, 3). Causes of sporadic PD are unknown but a prevalent hypothesis is that it may result from a complex interaction between toxic environmental factors, genetic susceptibility and aging (4).

Despite the fact that several gene expression studies focused on PD have been performed (5–14), no database has been developed to allow full-scale meta-analysis of microarray data related to PD.

Here, we present ParkDB, a resource that aims to provide comprehensive access to high quality gene expression datasets analyzed in a homogeneous manner. ParkDB has three main functions:

(i) To collect all microarray data from publicly available sources such as ArrayExpress (15), Gene Expression Omnibus (GEO) (16) and the Stanford microarray database (17), pertaining to PD and to dopaminergic neuron differentiation.
(ii) To provide access to key differentially expressed genes across different experiments contained in the database (i.e. different tissues, cell lines, treatments and species).
(iii) To provide homogeneous results from the statistical analysis of microarray data allowing for effective comparisons between experiments.

Database, Vol. 2011, Article ID bar007, doi:10.1093/database/bar007

Database schema and content

Data model

ParkDB is a relational database. The entity relationships between these tables are depicted in Figure 1. The conceptual scheme described was designed to facilitate the inclusion of new experiments and associated gene annotations in future updates.

ParkDB includes microarray data from human diseased brain areas and other tissues, as well as mouse, rat and zebrafish samples pertaining to PD affected tissues, PD animal models and PD-related experiments (Figure 2). Moreover, it also contains further annotation data associated to the genes identified and their orthologues. In total, more than 800 Affymetrix chips were re-analyzed.

Data insertion process

We download manually the raw data from GEO, ArrayExpress and Stanford databases every 2 months and re-analyze/re-annotate them using an automatic script written in R, which already contains annotation for all the Affymetrix platforms. Since ParkDB is built directly from raw data it is not affected by changing formats or data models of the databases we obtain the data from. Furthermore, researchers are invited to contact us to directly upload their data into ParkDB.

Queries

The ParkDB database allows the user to mine the data in the following ways:

(i) The 'gene query' can be performed by typing the NCBI gene name, the gene symbol, the Affymetrix probe identifier or the chromosome on which the gene is located. The resulting output view provides details on organism name, Affymetrix probe id, the description of each experiment where the gene was found to be differentially expressed at chosen cut-off, and the differential expression values

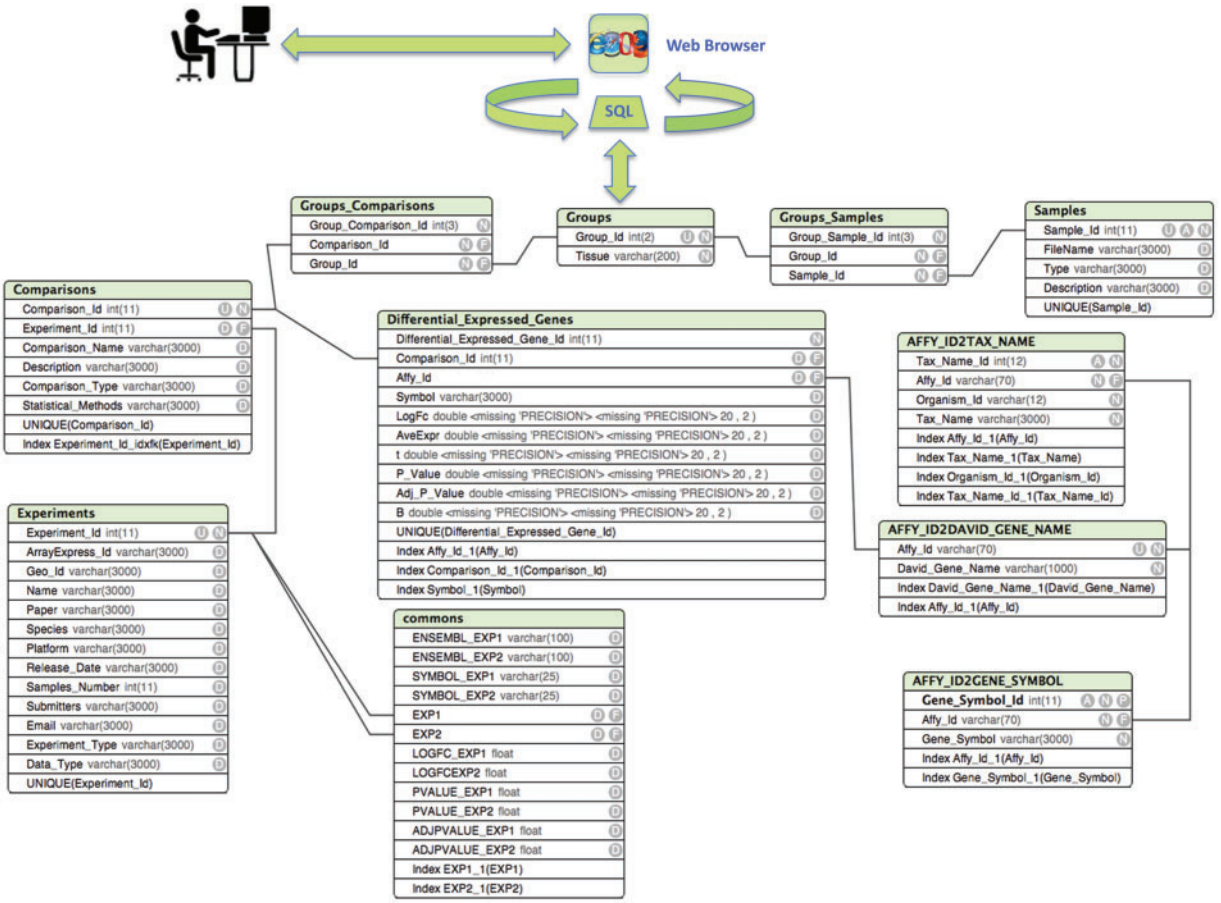

Figure 1. Entity relationship diagram of ParkDB. The scheme illustrates the tables created for data storage with the respective attributes and relations. The simple relations between the tables facilitate the possible expansion of the database to other genes involved in PD neurodegeneration and dopaminergic differentiation.

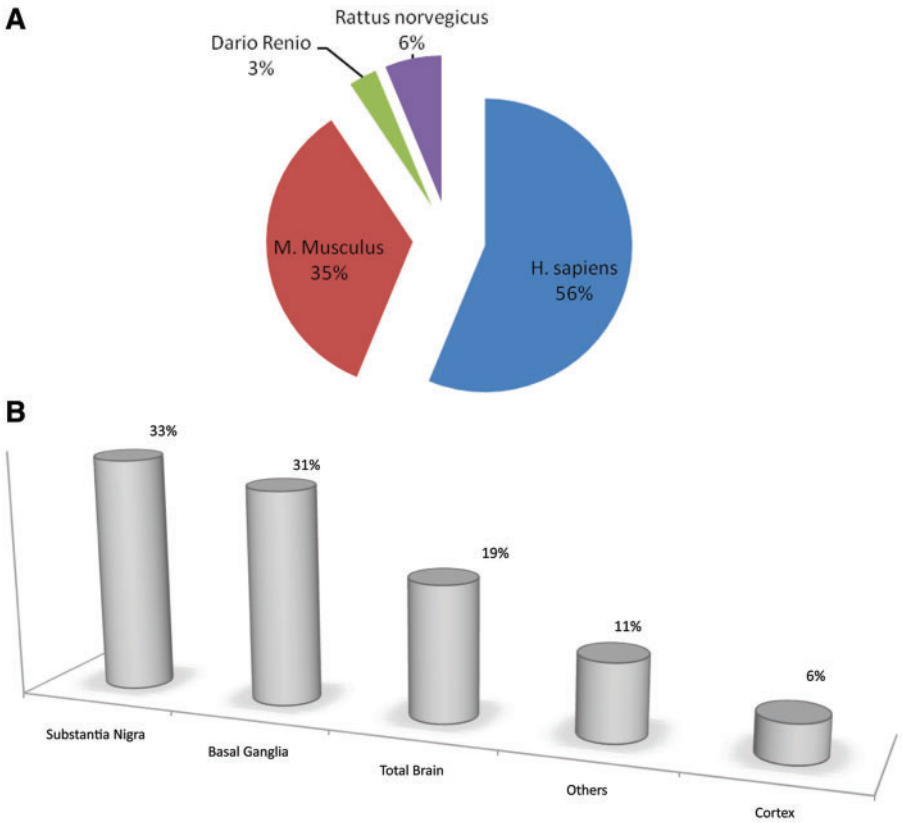

Figure 2. Pie charts and bar plot representing the curated information for each biological tissue, organism and comparison. Fifty-six percent of resources are retrieved from human experiments, 35% from mouse, 6% from rat and the remaining 3% from zebrafish. (**A**). Thirty-three percent of the tissues included in ParkDB were obtained from substantia nigra, 31% from basal ganglia, 19% from the whole brain, 6% from different regions of cortex and 11% from different tissues such as Blood, B Lymphocytes, etc. (**B**).

(i.e. log2 fold-change, *t*-test value, adjusted *P*-value, Bayesian value for *t*) (Figure 3).

(ii) The 'experiment query' allows searching for a specific experiment by names/keywords. In this case, a brief description table will be visualized alongside a list of all the genes, which were found to be differentially expressed (Supplementary Figure S1).

(iii) The 'tissue query' allows the researcher to search for specific tissue/cell types that may have been used in more than one experiment and retrieve the differentially expressed genes relating to those specific contrasts (Supplementary Figure S2).

(iv) The 'comparison query' provides a unique data-mining feature that has been made possible because of the homogeneous analysis of the expression data available in ParkDB. Since we have incorporated information on orthology, the researcher can query the database for common genes between two experiments and across different organisms. The resulting view will also highlight whether the genes were found to be up-regulated, down-regulated or anti-correlated, in two different comparisons (Figure 4).

Usage example

One of the most useful functions of ParkDB is the possibility to retrieve expression values for a particular gene in different experiments. For example, by querying ParkDB for gene symbols, we found that α-synuclein (SNCA) is differentially expressed and down-regulated in the entire set of human experiments comparing substantia nigra from Parkinson's patients against normal controls. SNCA is expected to be up-regulated in PD patients because it codes for a protein that forms Lewy bodies, which are the hallmark of PD.

This apparently inconsistent result can be explained by the fact that total mRNA was extracted from the remaining, more resistant neurons in post-mortem PD samples. Furthermore, Seo *et al*. (18) suggested a protective property for alpha-synuclein at low levels. They found that at nanomolar concentration alpha-synuclein protects rat cortical

Database, Vol. 2011, Article ID bar007, doi:10.1093/database/bar007

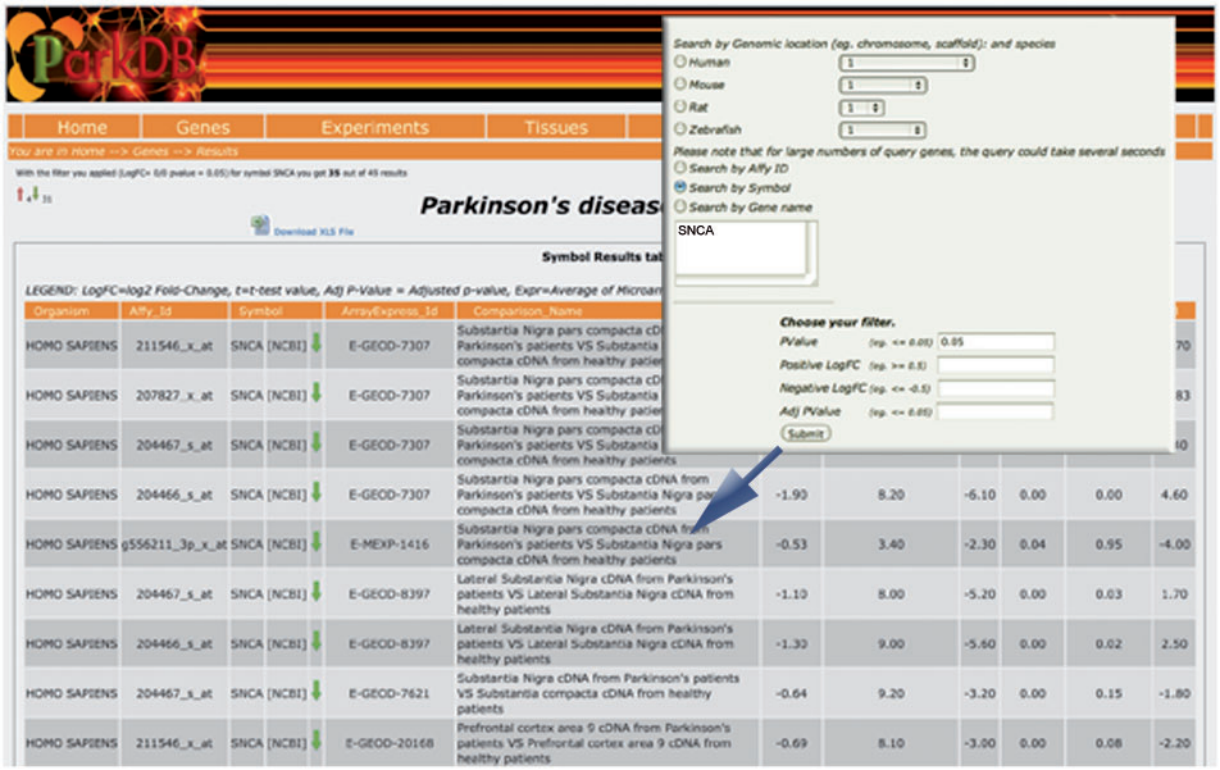

Figure 3. Gene query. The table shows the differential expressed genes obtained searching for SNCA. Red and green arrows indicate, respectively, up- and down-regulation.

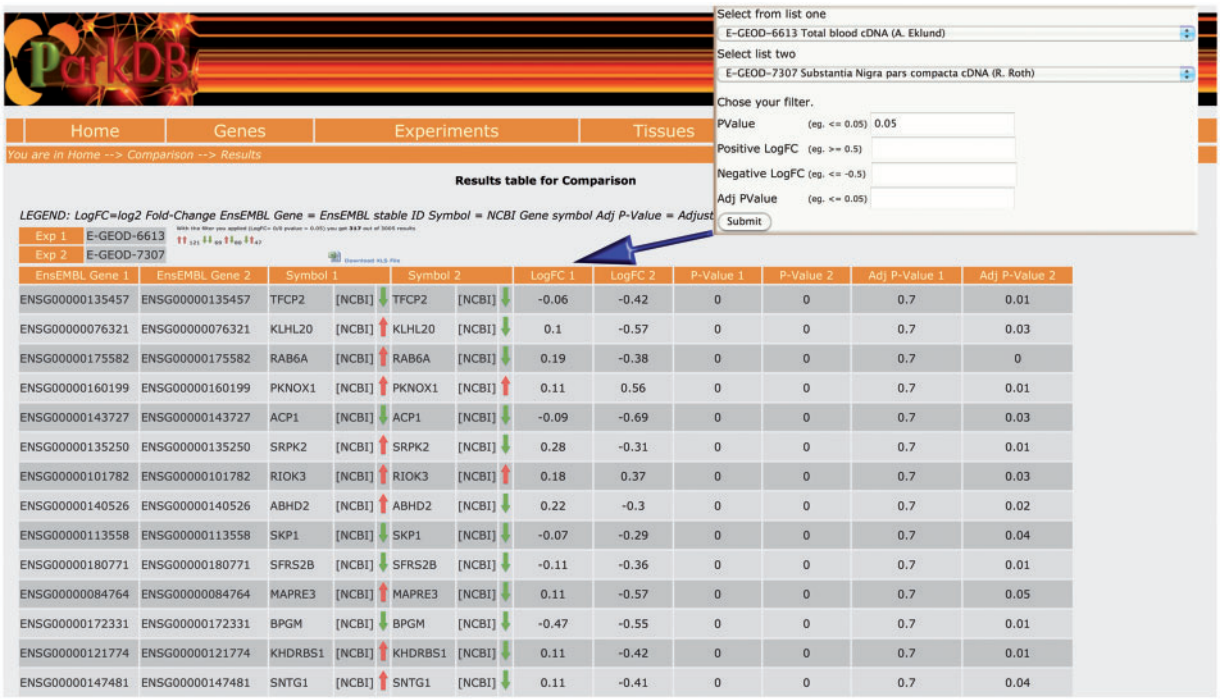

Figure 4. Comparison query. The table shows the common genes obtained comparing substantia nigra from PD patients and a transgenic mice that over-expresses HSP70/SNCA. The resulting view also highlights whether the genes were found to be up-regulated/down-regulated or anti-correlated, in two different comparisons.

Table 1. Gene expression table. SNCA (down-regulated), BCL2 (up-regulated) and AKT (up-regulated) are significant differentially expressed comparing substantia nigra in PD patients against normal control using ParkDB datasets

Gene name	Fold-change	Average expression value	*P*-value
SNCA	2.06 (down-regulated)	7.26	0.02
BCL2	1.82 (up-regulated)	6.25	0.01
AKT2	1.36 (up-regulated)	5.79	0.01

and hippocampal neuronal cells against serum deprivation and oxidative stress. This process is mediated by the Akt2 signaling pathway and its protective effect is increased by Bcl-2 overexpression. This study seems to confirm our results that show an up-regulation of BCL-2/AKT2 in human experiments comparing substantia nigra from Parkinson's patients against normal samples (Table 1) leading to the hypothesis that the anti-apoptotic effect of this pathway in the remaining cells of substantia nigra samples might represent a system that protects dopaminergic neurons from cell death. Moreover, higher BCL-2 levels in Alzheimer's disease are thought to be related to a compensatory up-regulation in response to the neurodegenerative process (19). Our findings might indicate that the expression profiling of post-mortem tissue from PD patients represents a specific signature of surviving neurons. The down-regulation of SNCA and other genes, however, might also be related to a significant limitation of post-mortem tissue, and/or reflect neuron loss and replacement by glia cells.

Data sources

We downloaded raw gene expression data from publicly available sources such as ArrayExpress, GEO and Stanford database. ArrayExpress is a repository of microarray data where files are both uploaded directly by researchers and imported from GEO. It is regularly queried for experiments based on keywords pertaining to PD and/or dopaminergic neuron differentiation. All the experiments which we import from ArrayExpress have passed their basic quality control (i.e. the raw data has to be available and no data files should be corrupted). In a few selected cases, we import directly data that is missing from Arrayexpress because it was uploaded on a project/publication specific website or obtained directly from the scientists involved. In that case, we perform this basic QC ourselves.

Using biomaRt (20) we were able to retrieve ortholog gene information from the Ensembl database (21) enabling users to perform inter-species queries. In order to allow among class comparisons (CC), we mapped all the CC genes using Ensembl Gene identifiers (EGIs). If an EGI was repeated in a single class comparison, we grouped them by considering the average, maximum, standard deviation of fold-change and the minimum adjusted *P*-value. Moreover, given two CCs, we provided the list of the genes measured in both the comparisons and, for each gene we provided the fold-change and the adjusted *P*-value. If both CCs were related to the same organism, we searched for those Ensembl identifiers that were included in both CCs, alternatively the ortholog gene identifier was used. Term annotations were derived from controlled vocabularies (22, 23).

Statistical methods

Analysis of all the microarray data incorporated in the database was performed using LIMMA (24), a Bioconductor (25) package for the R statistical programming language (http://www.r-project.org). The statistics employed by this package is based on the RMA method that provides quantile normalization, a linear transformation and a Bayesian *t*-test. The concept of using a *t*-statistic with a Bayesian adjusted denominator after a linear model transformation was first proposed by Smyth *et al*. (26). Each experiment was analyzed separately and no cross-experiment or cross-platform analyses were performed. This was done in order to avoid the limitations of source differences and batch effects. The comparison page allows the user to identify which genes have shown significant expression change across different experiments, but those values are derived from an analysis performed within each experiment separately.

Conclusions

ParkDB is the first database, which has collated, curated and re-analyzed the microarray data pertaining to samples from PD and dopaminergic neuron differentiation in a homogeneous manner.

It provides a unique combination of features that allow the researcher to perform data analysis and discovery of relationships between genes involved in PD in several human brain regions, as well as across multiple model organisms or cell lines.

Its strength lies in the careful re-analysis of all experiments, which ensures that the statistical basis for defining genes as differentially expressed is comparable. The goal of this study is to offer the researcher a central repository to perform meta-analysis queries.

By using the 'comparison' feature a researcher can easily download lists of genes, which are differentially expressed across tissues, cell lines, treatments and organisms, along with comparable and reliable statistical parameters for each gene in each experiment.

ParkDB will be regularly updated allowing the data to be refined continuously with every new ArrayExpress, GEO and Stanford microarray database version.

Further directions

Future improvements of ParkDB will allow the inclusion of next-generation sequencing data to fully exploit the relationship between microarray gene expression and deep-sequencing data in PD.

Supplementary Data

Supplementary data are available at *Database* Online.

Acknowledgements

The authors gratefully acknowledge the support and help of Jacek Marzek, Cristina Leonesi and all the members of DOPAMINET.

Funding

DOPAMINET project (www.dopaminet.org) founded by European Commission FP7 program (grant agreement number 223744).

Conflict of interest. None declared.

References

1. Samii,A., Nutt,J.G. and Ransom,B.R. (2004) Parkinson's disease. *Lancet*, **363**, 1783–1793.
2. Lesage,S. and Brice,A. (2009) Parkinson's disease: from monogenic forms to genetic susceptibility factors. *Hum. Mol. Genet.*, **18**, R48–R59.
3. Davie,C.A. (2008) A review of Parkinson's disease. *Br. Med. Bull.*, **86**, 109–127.
4. de Lau,L.M. and Breteler,M.M. (2006) Epidemiology of Parkinson's disease. *Lancet Neurol.*, **5**, 525–535.
5. Greco,D., Volpicelli,F., Di Lieto,A. *et al*. (2009) Comparison of gene expression profile in embryonic mesencephalon and neuronal primary cultures. *PLoS One*, **4**, e4977.
6. Chin,M.H., Qian,W.J., Wang,H. *et al*. (2008) Mitochondrial dysfunction, oxidative stress, and apoptosis revealed by proteomic and transcriptomic analyses of the striata in two mouse models of Parkinson's disease. *J. Proteome Res.*, **7**, 666–677.
7. Cantuti-Castelvetri,I., Keller-McGandy,C., Bouzou,B. *et al*. (2007) Effects of gender on nigral gene expression and parkinson disease. *Neurobiol. Dis.*, **26**, 606–614.
8. Yacoubian,T.A., Cantuti-Castelvetri,I., Bouzou,B. *et al*. (2008) Transcriptional dysregulation in a transgenic model of Parkinson disease. *Neurobiol. Dis.*, **29**, 515–528.
9. Lesnick,T.G., Papapetropoulos,S., Mash,D.C. *et al*. (2007) A genomic pathway approach to a complex disease: axon guidance and Parkinson disease. *PLoS Genet.*, **3**, e98.
10. Zhang,Y., James,M., Middleton,F.A. and Davis,R.L. (2005) 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.*, **137B**, 5–16.
11. Moran,L.B., Duke,D.C., Deprez,M. *et al*. (2006) Whole genome expression profiling of the medial and lateral substantia nigra in Parkinson's disease. *Neurogenetics*, **7**, 1–11.
12. Scherzer,C.R., Eklund,A.C., Morse,L.J. *et al*. (2007) Molecular markers of early Parkinson's disease based on gene expression in blood. *Proc. Natl Acad. Sci. USA*, **104**, 955–960.
13. Biagioli,M., Pinto,M., Cesselli,D. *et al*. (2009) Unexpected expression of alpha- and beta-globin in mesencephalic dopaminergic neurons and glial cells. *Proc. Natl Acad. Sci. USA*, **106**, 15454–15459.
14. Foti,R., Zucchelli,S., Biagioli,M. *et al*. (2010) Parkinson disease-associated DJ-1 is required for the expression of the glial cell line-derived neurotrophic factor receptor RET in human neuroblastoma cells. *J. Biol. Chem.*, **285**, 18565–18574.
15. Brazma,A., Parkinson,H., Sarkans,U. *et al*. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
16. Barrett,T. and Edgar,R. (2006) Mining microarray data at NCBI's Gene Expression Omnibus (GEO)*. *Methods Mol. Biol.*, **338**, 175–190.
17. Sherlock,G., Hernandez-Boussard,T., Kasarskis,A. *et al*. (2001) The Stanford Microarray Database. *Nucleic Acids Res.*, **29**, 152–155.
18. Seo,J.H., Rah,J.C., Choi,S.H. *et al*. (2002) Alpha-synuclein regulates neuronal survival via Bcl-2 family expression and PI3/Akt kinase pathway. *FASEB J.*, **16**, 1826–1828.
19. Satou,T., Cummings,B.J. and Cotman,C.W. (1995) Immunoreactivity for Bcl-2 protein within neurons in the Alzheimer's disease brain increases with disease severity. *Brain Res.*, **697**, 35–43.
20. Durinck,S., Moreau,Y., Kasprzyk,A. *et al*. (2005) BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, **21**, 3439–3440.
21. Hubbard,T., Barker,D., Birney,E. *et al*. (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
22. Kandel,E.R.E. (1981) Principles of neural science. Elsevier, North-Holland, NY.
23. Martin,E.A. (2007) Concise medical dictionary, 7th edn. Oxford University Press, Oxford.
24. Wettenhall,J.M. and Smyth,G.K. (2004) limmaGUI: a graphical user interface for linear modeling of microarray data. *Bioinformatics*, **20**, 3705–3706.
25. Zhang,J., Carey,V. and Gentleman,R. (2003) An extensible application for assembling annotation for genomic data. *Bioinformatics*, **19**, 155–156.
26. Smyth,G.K., Yang,Y.H. and Speed,T. (2003) Statistical issues in cDNA microarray data analysis. *Methods Mol. Biol.*, **224**, 111–136.