MDPD: an integrated genetic information resource for Parkinson's disease

Suisheng Tang^{1,*}, Zhuo Zhang¹, Gopalakrishnan Kavitha¹, Eng-King Tan² and See Kiong Ng¹

¹Data Mining Department, Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR) and ²Department of Neurology, Singapore General Hospital, Singapore

Received August 15, 2008; Revised September 18, 2008; Accepted October 7, 2008

ABSTRACT

Parkinson's disease (PD) is the second most common neurodegenerative disorder affecting millions of people. Both environmental and genetic factors play important roles in its causation and development. Genetic analysis has shown that over 100 genes are correlated with the etiology and pathology of PD. However, accessing genetic information in a consistent and fruitful way is not an easy task. The Mutation Database for Parkinson's Disease (MDPD) is designed to fulfill the need for information integration so that users can easily retrieve, inspect and enhance their knowledge on PD. The database contains 2391 entries on 202 genes extracted from 576 publications and manually examined by biomedical researchers. Each genetic substitution and the resulting impact are clearly labelled and linked to its primary reference. Every reported gene has a summary page that provides information on the variation impact, mutation type, the studied population, mutation position and reference collection. In addition, MDPD provides a unique functionality for users to compare the differences on the type of mutations among ethnic groups. As such, we hope that MDPD will serve as a valuable tool to bridge the gap between genetic analysis and clinical practice. MDPD is publicly accessible at http://datam.i2r.a-star.edu.sg/mdpd/.

INTRODUCTION

Parkinson's disease (PD) is a progressive neurological disease that affects millions of people world wide, with $\sim 1.8\%$ of the population at age >65 years (1). The death of dopaminergic neurons in the substantia nigra is a pathological feature of the disease. PD is a complicate disease that both environmental and genetic factors play various roles in its causation and development.

The genetic evidence is that first degree relatives of the familial PD patients are more vulnerable to PD than the general population, especially significant for early-onset PD (2,3). However, reliable biomarkers or tests to facilitate early and accurate diagnosis are currently not available (4). Genetic testing, if available, could complement clinical diagnostic criteria. Several genetic mutations and variants (point substitution, deletion, insertion or even polymorphisms) have been positively associated with PD (5,6). PARK2, LRRK2, PINK1, SNCA, UCHL1 and PARK7 are the most frequently studied genes (7). Some of the genetic variants are considered as causal factors while others may cause neuronal dysfunction indirectly. For example, mutations in the 5'-UTR of NR4A2 have significantly decreased the expression of NR4A2 gene and its downstream gene tyrosine hydroxylase (8). At present, over 100 genes have been reported to associate with PD in various forms. Some of these genetic variants are widespread in patients while others are ethnic-related risk factor. LRRK2 G2019S is a common pathogenic mutation found in 5-7% of familial PD and 1-2% of sporadic PD worldwide (9,10). At the same time, this mutation shows specific ethnic prevalence with exceptionally high frequency in North African Arabs (37-42% in familial and 41% in sporadic PD) (11) but is rare in Chinese (12,13). Another population-specific example is the GBA gene mutation. Both R496H and c.84insGGfs are found in patients from Ashkenazi Jews (14,15) and have not been reported in other ethnic groups. Genetic screening and treatment strategies could be improved if genetic features have been well characterized. Further elucidation of such information may also lead to new developments in diagnostic methods and early treatment.

Due to advances in genetic technology and the polygenic nature of PD, genetic information of the disease has accumulated rapidly in the past decade. The amount of data is a daunting challenge for individual researchers in searching and examining desirable information. For instance, there are over 1300 reports in the PubMed database if 'Parkinson's disease' AND 'mutation' are the keywords of search. The combination of gene names, official symbols

^{*}To whom correspondence should be addressed. Tel: +65 6408 2162; Fax: +65 6467 7152; Email: suisheng@i2r.a-star.edu.sg

^{© 2008} The Author(s)

and aliases in literature reports further amplifies the difficulty of retrieving relevant information. In addition, although information of gene function, gene sequences, protein structure and mutation reports are searchable, their availability are scattered across various databases such as Entrez Gene (16), GenBank (17), Swiss-Prot (18), OMIM (19) and PubMed (http://www.ncbi.nlm.nih.gov/ sites/entrez?db = PubMed). The Human Gene Mutation Database [HGMD; (20)], which covers mutation information for over 2800 human genes, is the most comprehensive but its public free-access version is limited to less-updated information. Moreover, it is not built specifically for PD. If one searches the HGMD for PD, one may only get the mutation information for less than 20 genes. Thus, to obtain desired information, researchers have to perform at least two time-consuming tasks: (i) examine large volume of data; and (ii) query a number of different databases. There are a few databases that specialize in PD mutation information, but they are limited in either coverage or functionality. For example, the LOVD Parkinson's disease mutation database developed by the Parkinson's Institute Leiden University (http://www.grenada.lumc.nl/ LOVD2/TPI/home.php) briefly covers six genes with total of 71 variants. The PDGENE (http://www.pdgene. org/) is more comprehensive and contains the most updated list of PD candidate genes with emphasis in genetic association studies. However, it lacks the functionality for ethnic comparison and it does not provide summary and statistic reports. To address some of the current limitations of PD databases and to facilitate effective and comprehensive information acquisition, we have developed the Mutation Database for PD (MDPD). Through a single online location with user-friendly interfaces, researchers are able to retrieve the latest information on PD, covering genetic variation (mutation and polymorphism), population studies, literature evidence and gene sequences. Various cross-references to public databases are incorporated to assist further exploration and evaluation.

DATA SOURCE AND CLASSIFICATION

MDPD is a specialist database that presents the human mutation information relevant to PD as an online resource. Data from animal models and cell lines are not included. Mutation evidences are extracted from PubMed database (from 1995 to 10 June 2008) based on keyword search ['Parkinson's disease/genetics' (MeSH)]. Sequences, variants and general gene information are obtained from other databanks as mentioned above (the latest update of MDPD was 4 September 2008). Every mutation entry has been manually examined by a researcher specialized in genetics. Important information, such as size of study sample, control group, population, age of onset, type of PD (sporadic versus familial), mutation outcome, its reference sequence and possible impact of the variant, have been manually extracted from the data source.

In addition to mutation, single nucleotide polymorphism (SNP) shows the linkage between genotype and the susceptibility of disease (21). SNPs can fall within the coding region and non-coding region in most cases.

In non-coding region, a SNP does not change protein sequence but may still have consequences on the risk of disease by affecting the splicing site or transcription factor binding site. Some SNPs have protective effects while others may increase the risk of PD, or show no significant outcomes and need further research. As we believe SNP information plays an important role in the design of genetic tests and also in understanding the mechanisms of the disease, they are included in our database. On the other hand, variations that lack precise genetic locations, such as variants in approximate chromosome regions or markers in inter-gene regions, are not included in MDPD. We believe genetic location is very critical in mutation study and such approximate results need to be resolved before inclusion in the database.

Case-control studies and genome-wide association studies (GWAS) have shown that PD is a polygenic disease that multiple gene mutations are responsible for the malfunction (7,22). Even monogenetic causes of PD could have resulted from a variety of mutations. The mutations in a particular gene can yield diverse consequences. A single nucleic acid substitution may lead to a multitude of possible outcomes, including amino acid exchange (missense mutation), no amino acid change (silent mutation), peptide truncation (nonsense mutation), absence of the protein (deletion) and even the production of a different protein (frame shift or insertion). Other types of mutation, such as duplication and compound mutation (more than one type of mutation) are also found in PD. As such, we categorize the genetic variations into the following: missense mutation, silent mutation, nonsense mutation, compound mutation, deletion, insertion, duplication, triplication, frame shift, short repeat and SNP. All of these classifications are based on the primary reference as the trusted data resource. We believe that such categorization can help the user to examine and compare the variations more efficiently. Another classification in MDPD is the variations' impact to reflect the outcomes of the variations. Divergent impacts are expected due to differences in the method used, sample size and the studied population. One variation may have more than one 'Impact'. For example, the V380L substitution in PARK2 gene is marked 'Associated' since it has been found in the sample of early onset PD patient (23). It is also tagged as 'Negative Result' in other occasion due to the lack of significant association with patients (24). Such inevitable discrepancy of impact reflects the complicated nature of PD in which multiple genetic factors play various roles and that interactions between genetic and environmental factors may influence the end result dynamically. According to the found effects of each variation, we classified its 'Impact' as 'protective factor', 'risk factor', 'associated', 'questionable' and 'negative result'. For example, if 'protective' or 'risk' effect has been mentioned in the primary reference of a variance, we assign its impact as 'protective factor' or 'risk factor'. 'Associated' is allocated to variances showing significant difference between patients and controls. If a variance has been reported in both patients and controls without statistic difference, we label it as 'questionable'. The classification aims to help user recapitulate information according to comparison outcome.

DATABASE STRUCTURE AND USAGE

MDPD is designed to be a publicly accessible online resource with user-friendly interface. MySQL, a reliable and proven relational database, is used to organize information. A web-based user interface to the database is provided via an Apache 2.0 HTTP server with PHP scripting engine.

MDPD contains the following functional pages: Browser, Search, Compare, Statistics and Variation Report. Three searching options are available in the search page: a variation search can be based on the gene name, gene ID or SWISS-PROT accession number. In accordance to the recommendations from the HUGO Nomenclature Committee, official gene symbols are used in each record, but MDPD also provides all known aliases in the result pages for easy reference and retrieval. For example, both PARK1 and SNCA are valid search terms in the database. Searching mutation information based on geographic region or the author's name in reference collection are two other helpful options. Searching for mutation information based on a geographic region enables a user to quickly know what genes have been studied in the region (or population) and the number of related publications may indicate regional research efforts. On the other hand, searching for mutation information based on the reference author's name can help researchers in the community easily identify various leading authors' research interests and their collaborators. Hyperlinks to each reference and gene symbol are provided in the report page.

Various useful features have been built into MDPD and span several web pages for ease of use and navigation. The complete list of web pages and their corresponding features are listed in Table 1. We highlight the 'Variation Report' page here as we believe that it is a useful feature of MDPD. In addition to providing the generic gene information, hyperlinks to Entrez, Swissport and OMIM, the 'Variation Report' page also covers detailed information of 'Variation Impact', 'Variation Type', 'Studied countries', 'Variation sequence' (in both amino acid and nucleic acid levels) and 'PubMed collection'. For 'Variation Impact' and 'Variation Type', we do not attempt to modify the findings from the primary reference, as we have mentioned previously. Users are advised to judge the classification based on his/her knowledge and the up-to-date research. To understand the impact of a genetic substitution, the user should be aware of any conflicting results from divergent ethnic groups or from specific subsets of patients. Such information is readily accessible in MDPD. 'Studied countries' provides information about the geographic regions and ethnic groups of patients studied. This information is also valuable for refining population screening target to avoid wasting resource. Researchers may also use it to identify key genetic factors, namely those for which impactful variations have been reported in many geographic regions (or ethnic groups). For example, MDPD includes 74 literature reports that described 258 different variants of PARK2 from 33 geographic regions. Deletion, duplication, triplication, insertion, missense mutation, nonsense mutation, silent

Table 1. Functional summary of MDPD

Web page	Contents		
Browse	Alphabetic list of gene symbol Chromosomal location of the gene Links to Entrez Gene and SWISS-PROT database		
Search	Entry to summary page Search gene name, symbol, aliases (allow partial name), gene ID, SWISS-PROT ID Search studied geographic regions (ethnic group)		
Summary	Search author's name for primary reference About the gene		
	Number of records for the gene Number of variants reported in SWISS-PROT database Link to OMIM database		
	Number of Pubmed reference for the gene in MDPD		
Variation report	List of variation impact List of variation type List of studied geographic regions (ethnic groups) Variation sequence (in both amino acid and nucleic acid levels) List of PubMed reference for the gene		
	Entry to individual variation report (sample size, control group, age, gender, testing variation, impact, geographic and comments)		
Comparison	Comparing genetic data from any two geographic regions		
Statistics	Key statistics in MDPD Top 10 genes with most literature reports Top 10 genes with most reported negative varian Top 10 countries/regions with most studies done		

mutation and compound mutations were all found in this gene. Among them, missense mutation and deletion are the most frequently conveyed with 140 and 86 records, respectively. Based on such information, a user can make reasonable inference about the importance of PARK2 instantly. To further confirm his/her speculation, the user can examine the primary reference through 'PubMed collection' and investigate the mutation 'hot sport' through 'Variation sequence'.

Allowing users to compare mutations between ethnic groups is another helpful element of MDPD. Users can readily obtain a list of mutation genes of an interested ethnic group from 'Search'. Comparing the mutations between two ethnic groups of interest can also be done easily in 'Comparison'. Currently, more than 2300 entries covering 202 human genes are stored in MDPD. Through systematic data mining on the integrated information, MDPD offers researchers new means for inspecting and making sense of the mutation evidences in published findings.

MDPD is publically accessible at http://datam.i2r. a-star.edu.sg/mdpd/.

DISCUSSION

Gene mutations and variations have become the focus of PD research in the last decade. Linkage mapping, case control study, pedigree analysis and GWAS are powerful approaches to identify and correlate genetic contribution

Table 2. Top 10 genes with the most published reference in MDPD

Gene Name	No. of record	No. of countries/regions	No. of reference
		, e	
LRRK2	265	38	76
PARK2	329	28	66
SNCA	117	24	51
CYP2D6	61	16	31
MAPT	39	14	20
PINK1	78	18	19
UCHL1	48	12	18
PARK7	58	12	16
MAOB	37	7	15
APOE	21	8	14

to PD. However, how various mutations and variants affect the disease and shape its development remains unclear. In addition, each mentioned approach has certain limitations. Various research biases and errors contribute to fewer reproducible association findings and diminish the assessing power between genetic variants and the risk of common disease (25,26). To partially overcome the limitation of accumulated imperfect data, we intend to include all published literatures with precise number of sample size (both case and control), variation position, variation impact and geographic location. We expect multiple independent genetic studies could yield meaningful results. At the same time, we remind users to be caution in interpretation of deductions from published data.

Many genes and multiple variants in a gene are registered positive correlation with PD. Possessing the available information is an essential first step to further understand and eventually to elaborate effective strategies for diagnosis and treatment. MDPD is an integrated information system that aims to facilitate PD research. It contains records for over 100 PD-associated genes verified from various genetic tests. Among them, the top 10 most reported mutation genes are LRRK2, PARK2, SNCA, CYP2D6, MAPT, PINK1, UCHL1, PARK7, MAOB and APOE (Table 2). The data in MDPD also reveal that current research has been focused in certain key genetic targets—the top 10 genes accounted for 1053 entries from 326 publications, which makes up to 44% of the total records and more than half of the literature reports (57%) in MDPD. At the same time, we realize that 202 genes are somewhat under the current research

Another interesting outcome from MDPD is the high frequency of 'negative result' in the variation reports. For example, >30% of records are labelled 'negative' in 9 out of the top 10 most reported genes (the exception is PARK2 with 15.2% negative reports). There are at least three implications: (i) many variants have low incident rates in PD patients and may not be a good screening target for survey; (ii) these variants may have insignificant impact to PD; and (iii) discrepancy may be caused by ethnic-related genetic variance, sample size, methods used or research errors. The variants with the most positive reports could be valuable genetic targets and further studies on them may warrant potential breakthrough in diagnosis and treatment.

PD is a multi-factorial disease for which the environmental factors and genetic elements are likely to be equally important. Studies have showed that lifestyles (such as smoking and coffee consumption), pesticides and metal exposure, and even well water drinking are factors that influence the risk of disease in both sporadic and earlyonset PD (27,28,29). The involvement of multiple genes, the high incident rate in aging population and high percentage of sporadic cases suggest the possibility of multiple interactions and connections in etiology of PD. In many cases, it is difficult to isolate the environmental factors and to specify the short- and long-term exposure. As such, we did not include environmental study information in MDPD, but the user should be mindful about the potential interactions with environmental factors.

FUTURE WORK

Discovering the relationships between the various genetic factors is an essential step toward understanding the mechanism of complex diseases such as PD. From MDPD, we know at least 202 genes have been examined for their possible involvement in PD. Our future work would be to develop an information system that can assess the impact of disease-causing mutations in terms of the functional changes of their encoded proteins and the interactions. Further integrated information, such as multiple level protein-protein interactions, and the role of the genetic variants in various neurodegenerative pathways will hopefully provide insights that will lead to novel treatments for PD.

ACKNOWLEDGEMENT

We thank Kar Leong Tew for amendment of the article.

FUNDING

Institute for Infocomm Research (I²R); Agency for Science, Technology and Research (A*STAR), Singapore.

Conflict of interest statement. None declared.

REFERENCES

- 1. de Rijk, M.C., Launer, L.J., Berger, K., Breteler, M.M.B., Dartigues, J-F., Baldereschi, M., Fratiglioni, L., Lobo, A., Martinez-Lage, J., Trenkwalder, C. et al. (2000) Prevalence of Parkinson. Neurology, 54, 21-23.
- 2. Clark, L.N., Ross, B.M., Wang, Y., Mejia-Santana, H., Harris, J., Louis, E.D., Cote, L.J., Andrews, H., Fahn, S., Waters, C. et al. (2007) Mutations in the glucocerebrosidase gene are associated with early-onset Parkinson disease. Neurology, 69, 1270-1277
- 3. Marder, K., Levy, G., Louis, E.D., Mejia-Santana, H., Cote, L., Andrews, H., Harris, J., Waters, C., Ford, B., Frucht, S. et al. (2003) Familial aggregation of early- and late-onset Parkinson's disease. Ann Neurol., 54, 507-513.
- 4. Chan, D.K. (2001) Parkinson disease and its differentials. Diagnoses made easy. Aust. Fam. Physician., 30, 1053-1056.

- 5. Dodson, M.W. and Guo, M. (2007) Pink 1, Parkin, DJ-1 and mitochondrial dysfunction in Parkinson. Curr. Opin. Neurobiol., 17, 331-337
- 6. Tan, E.K and Skipper, L.M. (2007) Pathogenic mutations in Parkinson disease. Hum. Mutat., 28, 641-653.
- 7. Farrer, M.J. (2006) Genetics of Parkinson disease: paradigm shifts and future prospects. Nat. Rev. Genet., 7, 306-318
- 8. Le, W.D, Xul, P., Jankovicl, J., Jiang, H., Appel, S.H., Smith, R.G. and Vassilatis, D.K. (2003) Mutations in NR4A2 associated with familial Parkinson disease. Nat. Genet., 33, 85-89.
- 9. Gilks, W.P., Abou-Sleiman, P.M., Gandhi, S., Jain, S., Singleton, A., Lees, A.J., Shaw, K., Bhatia, K.P., Bonifati, V., Quinn, N.P. et al. (2005) A common LRRK2 mutation in idiopathic Parkinson's disease. Lancet., 365, 415-416.
- 10. Kachergus, J., Mata, I.F., Hulihan, M., Taylor, J.P., Lincoln, S. Aasly, J., Gibson, J.M., Ross, O.A., Lynch, T., Wiley, J. et al. (2005) Identification of a novel LRRK2 mutation linked to autosomal dominant parkinsonism: evidence of a common founder across European populations. Am. J. Hum. Genet., 76, 672–680.
- 11. Benamer, H.T., de Silva, R., Siddiqui, K.A. and Grosset, D.G. (2008) Parkinson. Mov. Disord., 23, 1205–1210.
- 12. Lu, C.S., Simons, E.J., Wu-Chou, Y.H., Fonzo, A.D., Chang, H.C., Chen, R.S., Weng, Y.H., Rohé, C.F., Breedveld, G.J., Hattori, N. et al. (2005) The LRRK2 I2012T, G2019S, and I2020T mutations are rare in Taiwanese patients with sporadic Parkinson's disease. Parkinsonism Relat. Disord., 11, 521-522.
- 13. Skipper, L., Li, Y., Bonnard, C., Pavanni, R., Yih, Y., Chua, E., Sung, W.K., Tan, L., Wong, M.C., Tan, E.K. et al. (2005) Comprehensive evaluation of common genetic variation within LRRK2 reveals evidence for association with sporadic Parkinson's disease. Hum. Mol. Genet., 14, 3549-3556.
- 14. Aharon-Peretz, J., Rosenbaum, H. and Gershoni-Baruch, R. (2004) Mutations in the glucocerebrosidase gene and Parkinson. N. Engl. J. Med., 351, 1972-1927.
- 15. Gan-Or, Z., Giladi, N., Rozovski, U., Shifrin, C., Rosner, S., Gurevich, T., Bar-Shira, A. and Orr-Urtreger, A. (2008) Genotype-phenotype correlations between GBA mutations and Parkinson disease risk and onset. Neurology, 70,
- 16. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2007) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res., 35, 26 - 31.
- 17. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2008) GenBank. Nucleic Acids Res., 36, D25-D30.

- 18. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. and Bairoch, A. (2007) UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. Methods Mol. Biol., 406, 89-112.
- 19. McKusick, V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM, Am. J. Hum, Genet., 80, 588-604.
- 20. Stenson, P.D., Ball, E., Howells, K., Phillips, A., Mort, M. and Cooper, D.N. (2008) Human Gene Mutation Database: towards a comprehensive central mutation database. J. Med. Genet., 45, 124-126.
- 21. Naccarati, A., Pardini, B., Hemminki, K. and Vodicka, P. (2007) Sporadic colorectal cancer and individual susceptibility: a review of the association studies investigating the role of DNA repair genetic polymorphisms. Mutat. Res., 635, 118-145.
- 22. Maraganore, D.M., de Andrade, M., Lesnick, T.G., Strain, K.J., Farrer, M.J., Rocca, W.A., Pant, P.V., Frazer, K.A., Cox, D.R. and Ballinger, D.G. (2005) High-resolution whole-genome association study of Parkinson disease. Am. J. Hum. Genet., 77, 685-693.
- 23. West, A., Periquet, M., Lincoln, S., Lücking, C.B., Nicholl, D., Bonifati, V., Rawal, N., Gasser, T., Lohmann, E., Deleuze, J.F. et al. (2002) Complex relationship between Parkin mutations and Parkinson disease. Am. J. Med. Genet., 114, 584-591.
- 24. Oliveira, S.A., Scott, W.K., Nance, M.A., Watts, R.L., Hubble, J.P., Koller, W.C., Lyons, K.E., Pahwa, R., Stern, M.B., Hiner, B.C. et al. (2003) Association study of Parkin gene polymorphisms with idiopathic Parkinson disease. Arch. Neurol., 60, 975-980.
- 25. Ioannidis, J.P., Gwinn, M., Little, J., Higgins, J.P., Bernstein, J.L., Boffetta, P., Bondy, M., Bray, M.S., Brenchley, P.E., Buffler, P.A. et al. (2006) A road map for efficient and reliable human genome epidemiology. Nat. Genet., 38, 3-5.
- 26. Yesupriya, A., Evangelou, E., Kavvoura, F.K., Patsopoulos, N.A., Clyne, M., Walsh, M.C., Lin, B.K., Yu, W., Gwinn, M., Ioannidis, J.P. et al. (2008) Reporting of human genome epidemiology (HuGE) association studies: an empirical assessment. BMC Med. Res. Methodol., 8, 31.
- 27. McCulloch, C.C., Kay, D.M., Factor, S.A., Samii, A., Nutt, J.G., Higgins, D.S., Griffith, A., Roberts, J.W., Leis, B.C., Montimurro, J.S. et al. (2008) Exploring gene-environment interactions in Parkinson. Hum. Genet., 123, 257-265.
- 28. Elbaz, A. and Tranchant, C. (2007) Epidemiologic studies of environmental exposures in Parkinson. J. Neurol. Sci., 262, 37-44.
- 29. Aguiar Pde, C., Lessa, P.S., Godeiro, C. Jr., Barsottini, O., Felício, A.C., Borges, V., Silva, S.M., Saba, R.A., Ferraz, H.B. and Moreira-Filho, C.A. (2008) Genetic and environmental findings in early-onset Parkinson. Mov. Disord., 23, 228-233.