Published in final edited form as:

Clin Genet. 2011 August ; 80(2): 104–109. doi:10.1111/j.1399-0004.2011.01722.x.

Exome Sequencing in Parkinson's disease

Jose M Bras^{1,CA} and Andrew B Singleton²

¹Department of Molecular Neuroscience, Institute of Neurology, UCL, London, WC1N 3BG, UK

²Laboratory Of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, 20837, USA

Abstract

Exome Sequencing is rapidly becoming a fundamental tool for genetics and functional genomics laboratories. This methodology has enabled the discovery of novel pathogenic mutations causing mendelian diseases that had, until now, remained elusive. In this review we discuss not only how we envisage exome sequencing being applied to a complex disease, such as Parkinson's disease, but also what are the known caveats of this approach.

Keywords

Exome sequencing; Genetics; Parkinson's disease; Genomics

Introduction

Recent advances in genome technologies have allowed for a much broader understanding of the etiology of a vast number of disorders. These advances have thus far yielded significant results in two main fields: for common and complex diseases, genome-wide association studies (GWAS) have indicated genomic regions that confer risk for the disease; while for rare mendelian disorders, "second-generation" sequencing has pinpointed novel genes that contain mutations underlying the phenotype. A number of diseases, such as Parkinson's disease (PD), are suitable for both approaches: sporadic cases are amenable to GWA studies, while cases presenting a positive family history, strongly suggesting a mendelian form of the disease, are good candidates for sequencing-based studies. Between these two extremes of the frequency spectrum, are rare, but not causative, variants. These are not captured by GWAS but also do not have genotypic relative risks elevated enough to produce clear familial aggregation. Sequencing approaches will enable the study of these variants that, although carrying a lower relative risk than causative mutations, have higher risks than common variants, and thus, potentially higher impact in risk assessment, disease prevention and treatment.

PD is a progressive, incurable, neurodegenerative movement disorder that is clinically defined by bradykinesia, resting tremor and rigidity, and neuropathologically characterized by loss of neurons in the substantia nigra pars compacta and intraneuronal inclusions called Lewy bodies.

In recent years the understanding of the genetics of PD has seen great improvements: five genes are now known to be causative for monogenic forms (1-6), while eleven loci were

^{CA}Corresponding Author: Jose Bras, j.bras@ion.ucl.ac.uk .

The authors declare they have no conflict of interest to disclose.

recently identified as modulating risk for the development of common forms of PD(7, 8). Nonetheless, a significant proportion of inherited PD cases still remain unexplained genetically and the etiology of the disease remains, by and large, elusive.

The recent applications of exome sequencing have revealed a role for this approach for a number of diseases, in particular, but not exclusively, for mendelian diseases. Here we discuss how exome sequencing can be applied to PD research, and what are the caveats one needs to be aware of in this process.

Exome sequencing

Massively parallel sequencing was introduced approximately 3 years ago and it quickly became widely adopted by various laboratories working on genetics and functional genomics. This technology meant that, in theory, any laboratory could sequence entire human genomes in less than one month, a small fraction of the time taken to sequence the first human genome(9). However, this approach is not without its drawbacks: cost for a whole genome was still high, the infrastructure needed to handle these vast amounts of data was, in many cases, not available and interpretation of variants, particularly in the non-protein-coding portion of the genome was extremely challenging.

At this point, the need for a novel approach that would allow targeted sequencing of only a portion of the genome became apparent, and this led to the development of sequence capture/enrichment techniques, that have culminated in liquid-phase whole-exome capture kits that are currently widely used by the majority of groups interested in finding protein coding mutations that underlie disease.

Three major companies market these kits, and although there are small differences between them, the same principles remain: probes, complementary to all the known exons in the genome, are used to hybridize to single strand genomic DNA, after which they are magnetically separated to produce a DNA sample enriched for exons, this sample is then sequenced using second generation sequencing technology(10).

This approach allows the capture of relatively large amounts of non-sequential DNA; the current maximum is ~62Mb which is enough to contain the complete known set of coding exons in the genome (~30Mb) plus other regions prioritized as being of likely biological significance (e.g. microRNAs, promoters and UTRs). Currently these kits are able to capture > 97% of CCDS (<http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi>) and >96% of RefSeq coding exons (<http://www.ncbi.nlm.nih.gov/RefSeq/>), which is virtually the entire collection of known genes.

Although exome sequencing only allows for the study of a subset of the genome, making it clearly a transient technology, it has the advantages of lower cost, less time and easier interpretability, in comparison to whole-genome sequencing and analysis.

Exome sequencing as a research tool for PD

The search for genes underlying inherited forms of PD yielded remarkable results in the late 90s and early 2000s. Two genes, *SNCA* and *LRRK2*, were identified where mutations cause dominant forms of PD(2, 5, 6), while mutations in *PARK2*, *PINK1* and *DJ-1* were shown to underlie recessive forms of the disease(1, 3, 4). All five of these genes were identified using genetic linkage approaches, which require large pedigrees with affected and unaffected individuals. Following identification of sections of the genome that segregate with disease in each pedigree, sequencing is performed in genes located in those regions to identify the causal variant. Depending on the number of informative samples in the kindred, the linkage

region may contain modest to very large numbers of genes, preventing affordable Sanger sequencing of these genes. This process has some obvious drawbacks: it requires large numbers of individuals per family, it is costly, given that a large number of genes need to be sequenced by Sanger sequencing and it is time consuming since identifying the individuals and performing sequencing is a laborious process.

The most recent gene for PD to be identified by such methods was *LRRK2* in 2004 and, since then, no other genes have been found. Although this could suggest that no other PD genes exist, it is far more likely that the reason relates to the difficulty in obtaining informative pedigrees.

It has recently been shown that exome sequencing is an effective way to discover causative variants and genes. Ng and colleagues used this approach to sequence 12 human exomes(11). The study included four unrelated individuals with Freeman-Sheldon syndrome, a dominantly inherited rare Mendelian disorder. The investigators were able to identify variants in the known causative gene in each sample, which served as a proof-of-principle study. Since then, several disorders have been studied and novel genes identified, including Miller syndrome(12), Kabuki syndrome(13), severe brain malformations(14), hyperphosphatasia mental retardation syndrome(15), amyotrophic lateral sclerosis(16), among others.

This same approach is already being applied to PD and novel genes are certain to be identified. For clear recessive forms, pedigrees with as little as three or four individuals have potentially enough power to detect a novel pathogenic mutation(17). Likewise, dominant kindreds with only 4-5 individuals can be informative(18).

The majority of the studies so far have used public databases such as dbSNP(19) and the 1000Genomes Project(20) to filter the large amount of variants identified in their subjects (Fig. 1). The rationale for this approach is that for, rare diseases, even if they are recessive, the mutation will not occur in the general population at an appreciable frequency, thus such alleles will generally not be present in these databases. While this is true for rare disorders, it may not hold true for diseases such as PD, which have a significant prevalence in the population and high carrier frequency (e.g. heterozygous variants may be present in dbSNP, while the same variants when homozygous cause disease), in these cases, caution should be taken when filtering these results.

Another caveat of using exome-sequencing on a disorder like PD relates to the fact that some genes may have a high degree of sequence similarity with other genomic regions, including pseudogenes. The gene *GBA* is such an example; a pseudogene with over 96% sequence similarity is located just a few kilobases downstream of *GBA*, and it contains, in its normal sequence, nucleotides that are known pathogenic mutations when in *GBA*(21). Sequence reads that stemmed from the pseudogene may, in some instances, be aligned to the gene originating false positive mutations. This limitation is mainly a result of the relatively short read length of second-generation sequencing. Similarly, genes with high degree of repetitive sequences are also difficult to obtain information from.

These events, together with the fact that a small proportion of the known genes in the human genome is not being captured, are perhaps the major obstacles in the identification of novel gene mutations in properly defined familial samples.

In addition to studying kindreds, exome sequencing allows for association studies that are, to some extent, much more detailed than current GWAS using genotyping arrays. As sequencing costs continue to decrease, it becomes feasible to exome sequence large cohorts of cases and controls, allowing for association studies of the entire coding portion of the

genome. This will, in principle, reveal if any single gene in the genome plays a role in the phenotype; this resolution is something GWAS are unable to achieve. Similarly, GWAS were designed to assess only common variants; exome sequencing in large cohorts will enable the analysis of rare variants that are more likely to be functional(22). Again, *GBA* serves as an example: individually the PD risk conferring variants at *GBA* have low frequency (~1%) in the healthy population and were thus not detected within PD GWAS, even of relatively large sample sizes(7, 8). These variants however, have a relatively large effect on risk for PD, and would consequently be identified in a sequencing based association, provided that enough data was produced to overcome low quality mapping reads and that a large enough number of samples was tested. *GBA* mutations are calculated to have an odds ratio of approximately 5 and a combined frequency of ~0.02 in the control population, thus a study comprising ~1,000 case and 1,000 control samples would have >80% power to detect this association (using a multiple test correction for 21,000 genes in the genome). The identification of such variants, perhaps with even lower frequencies and higher relative risk, will allow for a more complete understanding of disease etiology. In addition, given that GWAS have yielded such a large number of loci for PD (currently eleven), it is expected that association studies using exome sequencing will not only help to define the reported associations by identifying novel mutations within genes at these loci(23) but also identify novel genes outside of these identified regions.

Perhaps the most efficient approach to exome sequencing in complex disorders with familial aggregation is to combine it with either linkage analysis(24), for dominant form of inheritance, or with homozygosity mapping(25), for recessive forms. Again, this requires moderate size, informative pedigrees, but has the benefit of simplifying filtering of variants to an enormous extent.

Hence, exome sequencing clearly has applicability in the study of Parkinson's disease, and both rare monogenic as well as common sporadic forms of disease are suitable to be studied using this approach. These data will undoubtedly identify novel genes harboring pathogenic mutations, as well as genes where risk variants play a role in the development of PD. Both aspects will further increase our knowledge of the etiology of disease.

Exome sequencing as a diagnostic tool for PD

Exome sequencing has already been shown to be a useful tool for diagnostic applications(26-29). The genetic diagnosis of congenital chloride diarrhea in a patient was recently made through exome sequencing revealing a homozygous missense variant in *SLC26A3*(30). Others have opted to use whole-genome sequencing and found that identical results would have been achieved by exome sequencing(31, 32), with the concomitant reduction in cost and increase in data interpretability.

To date there is no diagnostic test that can confirm PD. Diagnosis is usually made by clinical observation and confirmed only post-mortem by neuropathological studies. There are several implications to this, and perhaps the two most significant are the occurrence of misdiagnosis, that have been estimated to be as high as 25%(33), and the inability to perform pre-symptomatic diagnosis and pre-natal testing.

Genetic testing for PD is commonly directed only at the most likely gene(s) given the patient's clinical presentation. However, since PD is not only a genetic, but also a phenotypically heterogeneous disorder, this approach is likely to miss a significant proportion of genetic causes of disease. An example was recently shown in another neurodegenerative disease - ALS - where mutations in *VCP* (a gene known to harbor mutations causing Inclusion Body Myopathy, Paget disease and Frontotemporal Dementia), are also the cause of a small percentage of ALS cases(16).

For a proportion of PD patients exome sequencing can potentially be used as a screening method to identify pathogenic mutations.

When placed in the context of performing Sanger sequencing for all the known parkinsonism and related disorders' genes, exome-sequencing is rapid and cost effective. In addition, all these genes are thoroughly annotated and captured by current exome enrichment methods, further confirming that exome sequencing is a viable alternative for diagnostic screening of PD patients.

For PD, the majority of known pathogenic mutations are either point mutations or small deletions, which are both properly assayed using "second-generation" sequencing (Fig.2), unlike Spinocerebellar Ataxias or Huntington's disease where repeat mutations are the most common. Nonetheless, there are some caveats to the use of exome sequencing, as a diagnostic tool for PD. *PRKN* and *SNCA* are known to contain copy number mutations encompassing complete exons or even the entire genes (4, 34). Large copy number variants are difficult to detect using this approach. Similarly, large genomic rearrangements would also likely be missed. Read length is a further issue; even for sequencing chemistries allowing longer read lengths, phase is still impossible to determine for the vast majority of mutations, since these can potentially be as far apart as several megabases, and hence compound heterozygous mutations would not be immediately detected as such.

Conclusions and future prospects

Following on the success of GWAS, it is expected that, in the next few years, exome sequencing will reveal novel genes involved in myriad diseases. These will range from rare monogenic to common complex diseases, since this approach, unlike GWAS, is not based on the common variant, common disease hypothesis thus allowing the detection of rare alleles associated with disease, within the constraints of study design and power.

One of the major difficulties of exome sequencing is the amount of data generated, and the rapid evolution of methods to sift through these data. On average, any given exome will generate ~20,000 high quality variants, of which ~2,000 will be novel. This is clearly a very large number that needs to be filtered to identify the causal one(s), and until large cohorts of ethnically matched healthy individuals are sequenced on the same platforms, we will have to rely on public databases, which, although helpful, are not ideal for this purpose.

For monogenic disorders it is now clear that the best approach is to compare a proband's data with genomic information from the same family. This added information enhances the signal-to-noise ratio and allows for a dramatic reduction of the number of candidate genes. Under the existing framework, sequencing single affected individuals, for any given disease, is almost certain to not yield sufficient information to pinpoint novel causal variants.

For complex diseases, exome sequencing allows the execution of protein coding genetic association studies. Although these will certainly miss poorly annotated portions of the protein coding genome, whatever associations they return will definitely be easier to place in the context of disease involvement. These studies will, however, need large numbers of samples and are certainly dependent on the ongoing decrease in sequencing costs. Parkinson's disease will undoubtedly benefit from both of these approaches, and novel causes of, as well as novel risk factors for disease will be identified.

Although exome sequencing has shortcomings that are widely recognized its application has already led to remarkable discoveries, which are sure to be followed by many more as we continue to increase the number of samples and phenotypes being screened. Until whole-

genome sequencing becomes affordable in large-scale studies, exome sequencing has the potential to become the *de facto* screening method for a large number of diseases.

Acknowledgments

This study was supported in part by the Medical Research Council and Wellcome Trust disease centre and by the Intramural Research Program of the National Institute on Aging, National Institutes of Health, Department of Health and Human Services; project number Z01 AG000958-08

References

1. Valente EM, Abou-Sleiman PM, Caputo V, et al. Hereditary early-onset Parkinson's disease caused by mutations in PINK1. *Science*. 2004; 304:1158–1160. [PubMed: 15087508]
2. Paisan-Ruiz C, Jain S, Evans EW, et al. Cloning of the gene containing mutations that cause PARK8-linked Parkinson's disease. *Neuron*. 2004; 44:595–600. [PubMed: 15541308]
3. Bonifati V, Rizzu P, van Baren MJ, et al. Mutations in the DJ-1 gene associated with autosomal recessive early-onset parkinsonism. *Science*. 2003; 299:256–259. [PubMed: 12446870]
4. Kitada T, Asakawa S, Hattori N, et al. Mutations in the parkin gene cause autosomal recessive juvenile parkinsonism. *Nature*. 1998; 392:605–608. [PubMed: 9560156]
5. Polymeropoulos MH, Lavedan C, Leroy E, et al. Mutation in the alpha-synuclein gene identified in families with Parkinson's disease. *Science*. 1997; 276:2045–2047. [PubMed: 9197268]
6. Zimprich A, Biskup S, Leitner P, et al. Mutations in LRRK2 cause autosomal-dominant parkinsonism with pleomorphic pathology. *Neuron*. 2004; 44:601–607. [PubMed: 15541309]
7. Nalls MA, Plagnol V, Hernandez DG, et al. Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet*. 2011; 377:641–649. [PubMed: 21292315]
8. Plagnol V, Nalls MA, Bras J, et al. A two-stage meta-analysis identifies several new loci for Parkinson's disease. *PLoS Genet*. 2011
9. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 31:931–945.
10. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*. 2009; 27:182–189. [PubMed: 19182786]
11. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature*. 2009; 461:272–276. [PubMed: 19684571]
12. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies the cause of a mendelian disorder. *Nature genetics*. 2010; 42:30–35. [PubMed: 19915526]
13. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nature genetics*. 2010; 42:790–793. [PubMed: 20711175]
14. Bilguvar K, Ozturk AK, Louvi A, et al. Whole-exome sequencing identifies recessive WDR62 mutations in severe brain malformations. *Nature*. 2010; 467:207–210. [PubMed: 20729831]
15. Krawitz PM, Schweiger MR, Rodelsperger C, et al. Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nature genetics*. 2010; 42:827–829. [PubMed: 20802478]
16. Johnson JO, Mandrioli J, Benatar M, et al. Exome sequencing reveals VCP mutations as a cause of familial ALS. *Neuron*. 2010; 68:857–864. [PubMed: 21145000]
17. Glazov EA, Zankl A, Donskoi M, et al. Whole-Exome Re-Sequencing in a Family Quartet Identifies POP1 Mutations As the Cause of a Novel Skeletal Dysplasia. *PLoS genetics*. 2011; 7:e1002027. [PubMed: 21455487]
18. Wang JL, Yang X, Xia K, et al. TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. *Brain : a journal of neurology*. 2010; 133:3510–3518. [PubMed: 21106500]
19. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*. 2001; 29:308–311. [PubMed: 11125122]

20. Durbin RM, Abecasis GR, Altshuler DL, et al. A map of human genome variation from population-scale sequencing. *Nature*. 2010; 467:1061–1073. [PubMed: 20981092]
21. Hruska KS, LaMarca ME, Scott CR, et al. Gaucher disease: mutation and polymorphism spectrum in the glucocerebrosidase gene (GBA). *Human mutation*. 2008; 29:567–583. [PubMed: 18338393]
22. Gorlov IP, Gorlova OY, Sunyaev SR, et al. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet*. 2008; 82:100–112. [PubMed: 18179889]
23. Lehne B, Lewis CM, Schlitt T. Exome localization of complex disease association signals. *BMC Genomics*. 2011; 12:92. [PubMed: 21284873]
24. Southgate L, Machado RD, Snape KM, et al. Gain-of-Function Mutations of ARHGAP31, a Cdc42/Rac1 GTPase Regulator, Cause Syndromic Cutis Aplasia and Limb Anomalies. *Am J Hum Genet*. 2011; 88:574–585. [PubMed: 21565291]
25. Erlich Y, Edvardson S, Hodges E, et al. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res*. 2011; 21:658–664. [PubMed: 21487076]
26. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genet Med*. 2011; 13:255–262. [PubMed: 21173700]
27. Montenegro G, Powell E, Huang J, et al. Exome sequencing allows for rapid gene identification in a Charcot-Marie-Tooth family. *Annals of neurology*. 2011; 69:464–470. [PubMed: 21254193]
28. Bonnefond A, Durand E, Sand O, et al. Molecular diagnosis of neonatal diabetes mellitus using next-generation sequencing of the whole exome. *PLoS One*. 2010; 5:e13630. [PubMed: 21049026]
29. Johnson JO, Gibbs JR, Van Maldergem L, et al. Exome sequencing in Brown-Vialetto-van Laere syndrome. *Am J Hum Genet*. 2010; 87:567–569. author reply 569-570. [PubMed: 20920669]
30. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*. 2009; 106:19096–19101. [PubMed: 19861545]
31. Rios J, Stein E, Shendure J, et al. Identification by whole-genome resequencing of gene defect responsible for severe hypercholesterolemia. *Human molecular genetics*. 2010; 19:4313–4318. [PubMed: 20719861]
32. Lupski JR, Reid JG, Gonzaga-Jauregui C, et al. Whole-genome sequencing in a patient with Charcot-Marie-Tooth neuropathy. *The New England journal of medicine*. 2010; 362:1181–1191. [PubMed: 20220177]
33. Hughes AJ, Daniel SE, Kilford L, et al. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. *J Neurol Neurosurg Psychiatry*. 1992; 55:181–184. [PubMed: 1564476]
34. Singleton AB, Farrer M, Johnson J, et al. alpha-Synuclein locus triplication causes Parkinson's disease. *Science*. 2003; 302:841. [PubMed: 14593171]

Figure 1. General workflow used to identify the mutations underlying inherited forms of disease. This approach relies on the availability of both affected and unaffected individuals.

Figure 2. Exome sequencing results from a PD case showing a homozygous 1bp deletion in the gene *PARK2*. This mutation has been previously described as c.154delA and is shown in reverse complement.