Genetic Analysis in Neurology

The Next 10 Years

Alan Pittman, PhD; John Hardy, PhD

In recent years, neurogenetics research had made some remarkable advances owing to the advent of genotyping arrays and next-generation sequencing. These improvements to the technology have allowed us to determine the whole-genome structure and its variation and to examine its effect on phenotype in an unprecedented manner. The identification of rare disease-causing mutations has led to the identification of new biochemical pathways and has facilitated a greater understanding of the etiology of many neurological diseases. Furthermore, genome-wide association studies have provided information on how common genetic variability impacts on the risk for the development of various complex neurological diseases. Herein, we review how these technological advances have changed the approaches being used to study the genetic basis of neurological disease and how the research findings will be translated into clinical utility.


The diploid human genome is around 6 billion base pairs (bp) of DNA stored in 23 chromosome pairs. The Human Genome Project was initiated in 1990 to sequence the entire human genome from DNA from a number of anonymous individuals of predominantly European descent. The culmination of this work was the publication of the draft sequence in 2001, and by 2004, a high-quality reference sequence became available. Work by the Genome Reference Consortium continues to this day to improve the quality and coverage of low-complexity, repetitive, and hard to resolve regions.

Following on from the release of the reference genome, extensive analysis was performed to identify functionally significant regions. Although today the exact number of genes is still unknown, it is thought that there are approximately 21,000 protein-coding genes (1%-2%) contained in the human genome. The remainder of the genome consists of RNA genes, regulatory sequences, and repetitive DNA in which the function is poorly understood. However, recent work from the Encyclopedia of DNA Elements (ENCODE) project suggests that 80% of the human genome is indeed functionally active.1,2

Several different classes of DNA variation can occur between the genomes of different individuals. The most common type of variation is the single-nucleotide polymorphism. One would expect to find approximately 3 million such variants in any given individual compared with that of the reference sequence. These single base substitutions or point mutations arise, on average, every 1000 bp or so, and single-nucleotide polymorphisms that occur in more than 1% of the population are classified as common variants. These are often located in noncoding regions of the genome and tend to have little or no phenotypic effect. The vast majority of these common single-nucleotide polymorphisms have been extensively studied in many ethnically diverse populations by initiatives such as the International HapMap Project3 and constitute a valuable catalog and resource for genome-wide associa-
tion studies (GWASs) that investigate the effect of common variation on traits such as risk and susceptibility to common disease (e.g., type 2 diabetes mellitus and Alzheimer disease).

The single-nucleotide polymorphisms that occur in less than 1% of the population are classified as rare variants, and some of these may have profound phenotypic effects (e.g., such base changes can change or alter the sequence of a protein-coding gene). Genomic variation can also be caused by multiple base changes for insertion and deletion variants (i.e., insertions and deletions of bases that range in size from 1 to 1000 bp). Such variants can have a substantial effect in coding regions of the genome where they can result in gross alterations to a amino acid sequence or even a "frameshift" of the sequence resulting in a truncated protein. Larger insertions or deletions are referred to as copy number variants and can be both common and rare. Inversion and translocation events can also occur and can result in gross structural changes affecting many genes.

These types of variation can be present in germline cells or may be acquired somatically. Germline variation is either inherited directly or occurs de novo during meiosis or just after fertilization. Variation occurring in somatic cells is acquired and can arise randomly or through external environmental factors. Extensive somatic mutation is a hallmark of cancer but has also been implicated in autoimmune and neurodegenerative diseases.

Variation in DNA can also occur that contributes to heritable differences in gene expression; this is termed epigenetic vitiation. These modifications to the DNA include methylation and histone modifications, and they function without altering the DNA sequence itself and can change over time. Such modifications can have an important effect on disease (e.g., the switching off of tumor suppressor genes in cancer).

"NEXT GENERATION" OF DNA SEQUENCING

DNA sequencing in the laboratory has been possible since the 1970s, when the Sanger method was first developed, and has steadily improved and developed over time to facilitate automation and throughput. However, the technique remains too laborious and expensive (although >99.9% accurate) for the routine sequencing of whole genomes. Over the past 10 years, a number of new sequencing technologies have been developed that have significantly reduced the cost and time required for sequencing (Figure 1). These post-Sanger technologies are collectively described as next-generation sequencing (NGS) technologies and have been developed with whole-genome sequencing in mind. This, however, is not their sole purpose; they can be used for a wide range of applications, such as targeted resequencing and RNA sequencing.

Next-generation sequencing platforms have allowed for massive parallelization of sequencing reactions. Unlike the Sanger method in which each sequencing reaction represents a single predefined target, the DNA molecules in second-generation platforms are immobilized on a solid surface and are sequenced in situ. This allows for the sequencing of many millions of target molecules in parallel and for a substantial reduction in cost. Current NGS platforms use the clonal amplification of template DNA to generate "clusters" of identical DNA followed by sequencing through a stepwise incorporation of fluorescently labeled nucleotides or oligonucleotides. Since the middle of the last decade, there are 3 main commercial NGS platforms based on different sequencing chemistries. The technologies that are being used now in many laboratories are referred to as second-generation sequencing technologies to distinguish other technologies in the pipeline termed third-generation sequencing technologies.

Massive parallel sequencing has now allowed for an unprecedented interrogation of the variation in the human genome. For example, the 1000 Genomes Project, launched in January 2008, is an international collaborative research project involving the Wellcome Trust Sanger Institute (England), the Beijing Genomics Institute (China), and the National Human Genome Research Institute (United States), whose goal is to establish by far the most detailed catalog of human genetic variation. The plan is to sequence the genomes of 2500 anonymous participants from a number of different ethnic groups worldwide using a combination of methods: low-coverage genome sequencing and targeted resequencing of coding regions. The primary goals of this project are 3-fold: to discover single-nucleotide variants at frequencies of 1% or higher in diverse populations; to uncover variants down to frequencies of 0.1% to 0.5% in functional gene regions; and to reveal structural variants, such as copy number variants, insertions, and deletions. The results of a pilot project comparing different strategies for sequencing have already been published, and the sequencing of more than 1000 genomes was completed in May 2011. This resource is publicly available and can be used by researchers to identify variants in regions that are suspected of being associated with disease. By identifying and cataloging most of the common genetic variants in the populations studied, this project has generated data that will serve as an invaluable reference for clinical interpretation of genomic variation.

**THIRD-GENERATION SEQUENCING TECHNOLOGIES**

Massively parallel sequencing has become the dominant sequencing technology, but other approaches have

---

**Figure 1.** Evolution of DNA sequencing technologies (adapted from Stratton et al). PCR indicates polymerase chain reaction.
emerged that avoid amplification of the DNA template prior to sequencing and instead aim to sequence the single DNA molecule in real time. These new technologies are collectively referred to as the “next” NGS or third-generation sequencing (Figure 2). The potential benefits of using single-molecule sequencing are minimal input DNA requirements, elimination of amplification bias, faster turnaround times, and longer read lengths that allow for some haplotyping of sequence information.

ANALYSIS OF DATA AND THE INFORMATICS PIPELINE

The volume of data generated by NGS is enormous, and the workload has shifted away from the laboratory toward the data analysis process. Analysis of whole-exome or whole-genome data requires substantial computation, data storage, and informatics tools for interpreting the variant data (Figure 3). This is the least trivial aspect of NGS and represents the true challenge.

The analysis pipeline for NGS technology can be roughly divided into 3 analytical steps:

Primary analysis: base calling—converting the light-signal intensities to nucleotide base calls (usually done by the onboard machine software while running).

Secondary analysis: alignment and variant calling—mapping the short DNA reads to the reference sequence and calling differences or variants between the two.

Tertiary analysis: interpretation—analysis of the variant data with respect to the genetic experiment in question.

Although these are all fairly standardized processes now, there are a few points of note relating to the alignment and annotation of variants. Numerous programs have been developed specifically for alignment and variant calling; however, a major issue for standard alignment programs is the interpretation of small insertions and deletions, which has been partly addressed with new programs for this purpose. However, current technology and analysis do not allow for fully confident analysis of insertions and deletions. In the final stages of the alignment phase, the data are annotated with genetic and biological information that can be visually inspected through a graphical interface. The ability to provide accurate and comprehensive genome annotation is critical for interpretation and for the implementation of filtering steps to exclude nonpathogenic and irrelevant variants. Once the majority of variants have been excluded, a small number of potential variants may remain that could be linked to the disease phenotype.

GENETICS AND NEUROLOGICAL DISEASE

The genome of any given individual will contain millions of sequence variants of which the vast majority will have no effect (neutral variation) or will represent normal differences in phenotype (eg, hair color). However, some may harbor pathogenic mutations that cause or predispose to disease. Determining if a single variant is associated with a disease can be a slow process, especially if the effect is subtle.

Monogenic gene disorders are usually associated with rare, highly penetrant genetic mutations that have a profound effect on the function of a gene (eg, by changing the coding sequence). However, the severity and penetrance of the phenotype can vary widely, and this could
be due to the influence of other modifier genes. Such single-gene disorders tend to run in families with a clear inheritance pattern. In addition to rare, highly penetrant mutations, common variants in the population contribute to the susceptibility to common, complex neurological disease. These variants tend to have small effects on risk and are usually found in the noncoding portion of the genome. Assessing disease risk at the individual level based on these variants is challenging and, generally speaking, has limited clinical utility.

Parkinson disease (PD) is the second most common neurological disease of adult onset, with increased incidence with age; approximately 10% of patients report a family history. 

Mendelian forms of PD occur with both autosomal dominant and recessive patterns of inheritance. Toxic gain-of-function mutations in SNCA, LRRK2, and VPS35 cause autosomal dominant PD, and, furthermore, common polymorphisms in SNCA and LRRK2 exert a small but significant risk effect on non-Mendelian forms of PD. Conversely, loss-of-function mutations in PARK2, PARK7, and PINK1 cause autosomal recessive PD.

Collectively, the monogenic forms of PD account for about 30% of familial cases and approximately 5% of sporadic cases (Table 1). Although non-Mendelian forms of PD show a relatively low level of heritability, a larger number of susceptibility loci (including SNCA, MAPT, and LRRK2; Table 2) have recently been identified through GWASs, and although their exact mode of action has yet to be elucidated, it is most likely that susceptibility acts through subtle changes in the gene expression of the target genes. These susceptibility variants are considered of low penetrance and would be expected to decrease risk, on average, by 10%. Recessively transmitted GBA mutations cause Gaucher disease, a lysosomal storage disease, and relatives of patients with Gaucher disease show an increased incidence of PD. Subsequent genetic studies of GBA revealed that rare polymorphisms have a role to play (ie, significantly increasing the risk of PD 5-fold).

Many other neurodegenerative disorders show an extensive family history. For example, Alzheimer disease, frontotemporal dementia, and amyotrophic lateral sclerosis show rare but significant familial inheritance, Mendelian forms of diseases, and lower-penetrance variants associated with the more common sporadic forms of disease.

HIGH-THROUGHPUT GENETIC APPROACHES TO NEUROLOGICAL DISEASE RESEARCH: THE ROAD AHEAD

In recent years, neurogenetics research had made some remarkable advances owing to the advent of the genotyp-
ing arrays and NGS techniques herein described. These new techniques allow for increasingly larger numbers of samples of the genome to be interrogated at high resolution. These advances have come after 20 years of small candidate gene studies and traditional positional cloning techniques. Since the advent of GWASs in 2005, several well-replicated neurodegenerative studies have identified new disease loci, and these discoveries are set to continue over the next few years with ever-increasing sample sizes under study. Despite the achievements of GWASs, this approach is limited. First, it is only able to study relatively common types of variants, those that occur at a frequency of more than 1% in the general population. Second, the major problems associated with the discovery of such risk-associated variants is the interpretation of the risk in the context of disease pathogenesis.

The results from a typical GWAS highlight disease-associated regions of the genome that can be several kilobases or, indeed, even up to a megabase in size, and because such variants tend to be noncoding, it is not entirely obvious what the target gene or functional consequence is. One possible approach for resolving this issue would be to integrate high-throughput genotype data, sequencing data, and local gene expression data in such a way that the biological basis of the association can be elucidated.

Occasionally, the biological basis for such disease associations can be relatively straightforward to dissect (e.g., at the SNCA locus in PD). In rare cases, Mendelian mutations can cause disease by being toxic gain-of-function coding mutations, and (in parallel) common, noncoding variants in the population are present in the same gene that has a more modest effect on disease risk by subtly altering the regulation and expression of the SNCA gene. However, for the vast majority of other GWAS “hits,” further work is needed to dissect out the precise causal variants and their effect on disease.

One such approach is to use genome-wide expression quantitative trait loci (eQTLs) data sets. These eQTLs are genomic loci that regulate the expression levels of messenger RNA (mRNA). The measured mRNA is the product of a single gene with a specific chromosomal location. The eQTLs may act locally cis or trans (at a distance) of a gene, and the abundance of a gene transcript is directly modified by a polymorphism in a regulatory element. The combination of GWAS and the measurement of global gene expression allows for the systematic identification of eQTLs. By assaying gene expression and genetic variation simultaneously on a genome-wide basis for a large number of individuals, statistical methods can be used to map the genetic factors that underpin individual differences in the quantitative levels of expression of many thousands of transcripts. Such data sets are becoming widely available, for example, the UK Human Brain Expression Consortium data set, generated from 10 distinct brain regions sampled from 134 neuropathologically normal individuals, contains detailed information on the regional expression, splicing, and regulation of genes in physiologically relevant tissue. Already this approach has elucidated GWAS neurological disease hits (e.g., the association of the common 17q21.31 H1 MAPT locus with PD). Studies on eQTLs have revealed that this risk polymorphism is associated with an increased expression of exon 3 containing MAPT transcripts in the human brain. This biologically significant finding opens up new lines of research into the role of tau protein in PD and related neurodegenerative conditions.

Over the last 2 years, whole-exome sequencing has rapidly become the approach of choice to study rare variations that are not captured by GWASs. It is inexpensive but still an effective alternative to whole-genome sequencing, especially since approximately 85% of Mendelian disease-causing mutations are located within 1 of the 180,000 coding exons, which constitutes a mere 30 MB or 1% of the genome. Nonetheless, whole-exome sequencing is not without its limitations; for example, not all known genes are well captured or well sequenced using this technology. However, it is anticipated that, over the next few years, this approach will be gradually replaced by whole-genome sequencing owing to the decreasing cost of sequencing and the far greater variant information content gleaned from the entire genome.

Owing to recent advances in high-throughput genotyping and sequencing technologies, it is hoped that genetic research will uncover a large number of additional disease-causing and disease-modifying sequence variants over the coming years. There is no doubt that these new discoveries will lay the foundations for initiating new biological research and will open up avenues to new treatment approaches and diagnostics. The genome-wide approaches described herein have, without a doubt, advanced the field of neurogenetics. With the continuing evolution of and improvements to the technology, its rapidly decreasing cost, and the improved data informatics of NGS, it is anticipated that almost all routine genetics research will be conducted in this way.

GENETIC TESTING FOR NEUROLOGICAL DISEASE IN A CLINICAL SETTING

It is expected that this wealth of new information on the genome and its effect on the risk of neurological disease will result in the development of novel diagnostic assays and targeted therapies and in an improved ability to predict the onset, severity, and progression of disease. In other words, it will have a major impact on medical practice. Furthermore, the advances in NGS described herein that have transformed genomic research now have the potential to revolutionize the way in which genetic neurological diseases are diagnosed in the laboratory in a clinically useful way.

It is highly desirable to use NGS in the diagnosis of neurological disease, and many such new tests are beginning to become available and are being offered either commercially by local health trusts. The development of these new high-throughput methods is advantageous for several reasons. There is wide clinical and genetic heterogeneity of neurological diseases, which means that any given disorder may present with a wide spectrum of clinical phenotypes and that even mutations in the same gene may present with different syndromes. For example, mutations in the F2H gene may present in the clinic as brain iron accumulation, leukodystrophy, or hereditary spastic paraplegia. Clinical heterogeneity can also
arise from mutations in several genes associated with a clinically typical phenotype (e.g., LRRK2 and SNCA multiplications) and, likewise, in the recessive genes PARK2, PINK1, and PARK7. In addition, numerous genes can underlie a Huntington disease phenotype. Thus, it is difficult to assign a specific gene test just on clinical grounds alone, and a strong case can be made for multigene testing of a disease area rather than a single phenotype or a possible single underlying gene.

Many important points that can barely be touched on here need to be considered before NGS is routinely incorporated into the diagnostic laboratory, but they would have to include a careful consideration of (1) the selection of the appropriate technology, (2) the ethical and legal issues, and (3) the data analysis and the infrastructure of the information technology.

It is our view that, at this point in time, the most efficient and appropriate strategy is that of clinically targeted analysis, using one of the benchtop sequencers discussed previously, rather than whole-exome sequencing or whole-genome sequencing, such that only genes of relevance to the specific condition are analyzed and the results of which are shared with patients. That is not to say that such new tests will focus only on a single gene but, rather, that all possible known causal genes with the potential to cause a condition will be screened. For example, in the case of PD, a targeted test that screens all possible causal Mendelian genes could include the following typical set: SNCA, LRRK2, PARK2, PARK7, PINK1, and VP533. However, one would probably wish to include genes that have been associated with other atypical forms of parkinsonism such as ATP13A2, FBX07, and PLA2G6 and genes such as MAPT and GCH1 associated with phenocopies of PD. Thus, in this case, a specific diagnostic question is whether to “test” the genome sequence for known clinically validated pathogenic variants.

Whole-exome sequencing is unlikely to be routinely adopted by the diagnostic market owing to the limitations in the technology (i.e., poorly captured genes). The consequences of not capturing all the known genes can be illustrated by the GBA gene. As previously discussed, heterozygous mutations in GBA are the strongest genetic risk factor for developing PD to date.24 GBA poses difficulties for sequencing because it has a pseudogene with approximately 96% homology a few kilobases downstream. This poses difficulties for capturing the target DNA, for postsequence alignment, and for variant calling.

In the years to follow, it is anticipated that whole-genome sequencing will become a widespread tool for clinical use because of its decreasing cost, the new advances in the technology, and the improved informatics systems that can handle the large volume of data generated. This application of whole-genome sequencing will be particularly important in the diagnoses of the conditions of individuals for whom traditional sequencing approaches have failed to identify the underlying cause of disease. One such new initiative is the National Institutes of Health Undiagnosed Disease Program.25 In the first year of this program, 160 individuals were enrolled, and for 39 cases, a diagnosis was able to be made. Of the 160 individuals in the cohort, 85 (53%) had a neurological disorder that was successfully diagnosed using whole-genome sequencing.

Performing whole-genome sequencing has the potential to provide answers to the diagnostic questions about the medical condition and about the potential predispositions to unconsidered conditions in the future, which may have implications for other family members. Thus, the routine implementation of whole-genome sequencing will necessitate a detailed review of the full ethical implications for health care providers and for the patients themselves.

Accepted for Publication: December 6, 2012.
Published Online: April 9, 2013. doi:10.1001/jamaneurol.2013.2068

Correspondence: John Hardy, PhD, Department of Molecular Neuroscience and Reta Lila Weston Laboratories, Institute of Neurology, Queen Square House, University College London, 9th Floor, Queen Square, London WC1N 3BG, England.

Author Contributions: Study concept and design: All authors. Acquisition of data: Pittman. Analysis and interpretation of data: Pittman. Drafting of the manuscript: All authors. Critical revision of the manuscript for important intellectual content: Pittman. Statistical analysis: Pittman. Obtained funding: Hardy. Administrative, technical, and material support: Pittman.

Conflict of Interest Disclosures: None reported.

REFERENCES


