The Neurogenomics View of Neurological Diseases

Shoji Tsuji, MD, PhD

The availability of high-throughput genome sequencing technologies is expected to revolutionize our understanding of not only hereditary neurological diseases but also sporadic neurological diseases. The molecular bases of sporadic diseases, particularly those of sporadic neurodegenerative diseases, largely remain unknown. As potential molecular bases, various mechanisms can be considered, which include those underlying apparently sporadic neurological diseases with low-penetrant mutations in the gene for hereditary diseases, sporadic diseases with de novo mutations, and sporadic diseases with variations in disease-susceptible genes. With unprecedentedly robust power, high-throughput genome sequencing technologies will enable us to explore all of these possibilities. These new technologies will soon be applied in clinical practice. It will be a new era of datacentric clinical practice.


The elucidation of the molecular bases of neurological diseases is fundamental to the development of disease-modifying and preventive therapies. Over the past 3 decades, we have witnessed remarkable progress in the identification of the genes that cause hereditary neurological diseases (Figure 1). This has been accomplished mainly on the basis of the research paradigm known as “positional cloning,” which uses linkage studies to pinpoint the position of genes on chromosomes followed by the identification of the causative gene. The identification of causative genes has further made it possible to develop disease models for hereditary neurological diseases and to develop therapeutic strategies. The majority of neurological diseases, however, are sporadic without any obvious familial occurrence. We are thus faced with the challenge of elucidating the molecular bases of sporadic diseases. Intriguingly, the clinical presentations and neuropathological findings of hereditary forms of neurodegenerative diseases are often indistinguishable from those of sporadic diseases, raising the possibility that common pathophysiologic pathways underlie both hereditary and sporadic neurodegenerative diseases.

In contrast to the molecular bases of hereditary neurological diseases, the molecular bases of sporadic neurological diseases, particularly those of sporadic neurodegenerative diseases, largely remain unknown. A potential clue to the molecular bases of sporadic neurological diseases may be the clinical observation that siblings and relatives of a patient with a neurological disease are at an increased risk of developing the same disease; this phenomenon has been observed with regard to Parkinson disease (PD) and amyotrophic lateral sclerosis. These clinical observations suggest the involvement of genetic factors in these diseases (Figure 1). Until recently, it has been difficult to elucidate the genetic factors underlying sporadic neurological diseases.

CME available online at jamanetworkcme.com and questions on page 683

Rapid advancements in genome science, particularly the availability of massively parallel sequencing technologies that use next-generation sequencers (NGSs), are revolutionizing the neurogenomics view of
Emerging new technologies for nucleotide sequencing have brought about a remarkable revolution in analyses of the human genome sequence. Compared with a conventional technology (namely, the Sanger method), the throughput of massively parallel sequencing that uses NGSs is increasing dramatically, with the current throughput at 600 GB per run, which means that a sufficient amount of sequence data can be obtained for whole-genome sequencing of at least 4 individuals. In typical whole-genome sequencing experiments, billions of short reads (100-150 base pairs) are obtained. These short reads are aligned to human genome reference sequences, and sequence variations are called through computational analyses.

Currently, 2 types of sequencing strategy (namely, whole-exome and whole-genome sequence analyses) are used. Because the cost of whole-genome sequencing is still considerably high, it is not easy to conduct whole-genome sequencing for a large number of individuals. In whole-exome sequence analysis, the enrichment of exonic sequences using oligonucleotide “baits,” which is followed by sequencing, has been preferentially used. With this strategy, all exonic sequences in the human genome can be efficiently enriched. With this approach, more than 90% of target regions can be enriched, and these enriched genomic regions are then subjected to massively parallel sequence sequencing using NGSs. This approach is currently being used a lot for the identification of disease-relevant variants and even for diagnostic purposes.

Given the ever-increasing throughput of NGSs and the dramatically decreasing costs, it will soon be a realistic approach to conduct whole-genome sequencing for various research applications. Studies have shown that there are more than 3 million variations in the human genome of each individual. In one study, among the 3.3 million single-nucleotide polymorphisms (SNPs), 8996 known nonsynonymous SNPs and 1573 novel nonsynonymous SNPs were identified. Interestingly, 32 alleles exactly matched mutations previously registered in the Human Gene Mutation Database. In addition, 345 insertions/deletions were observed to overlap in a coding sequence and may alter protein function. These findings indicate that, among the numerous candidate variations, it will be a challenge to determine which variations are relevant to diseases.

Given the enormous number of short read sequences (~100 bp), informatics analyses, including mapping to reference sequences and indentifying variations, require a huge computational power. Furthermore, mutations can be variable, including single base substitutions, insertions/deletions, and structural variations. It is difficult to efficiently identify all the variations using currently available NGSs and software. For example, expansions of repeat motifs identified in frontotemporal dementia and amyotrophic lateral sclerosis are difficult to identify using NGSs.

As already stated, most of the currently available NGSs produce billions of short reads of 100 to 150 bp. This is the limitation in analyzing various structural variations, some of which may be relevant to neurological diseases. Very recently, single-molecule sequencing technology has become available from Pacific Biosciences; this type of technology enables the acquisition of nucleotide sequences as large as 10 kilobases. Another single-molecule sequencing technology using nanopores, which allows for the acquisition of much longer sequences, will soon become available.
individuals is often limited owing to small family sizes and the small number of family members with a confirmed clinical and/or a pathological diagnosis. These circumstances pose a challenge to positional cloning because the candidate regions cannot be narrowed down to small regions that are sufficient for identifying the causative genes by sequencing individual genes in the candidate regions. Despite these difficult circumstances, the availability of NGSs with unbelievably high throughput has made the identification of causative genes possible. Given the large capacity of NGSs, the most essential step (and the bottleneck) is now the collection of as many samples from patients and their families as possible based on well-characterized clinical information, including the correct diagnosis, regardless of family size or number.

**EFFECT OF HIGH-THROUGHPUT GENOME SEQUENCING ON UNDERSTANDING THE MOLECULAR BASES OF SPORADIC NEUROLOGICAL DISEASES**

The elucidation of the molecular bases of sporadic neurological diseases is now a big challenge. We need to take various mechanisms into account as the molecular bases of sporadic neurological diseases, which include (1) apparently sporadic diseases with low-penetrant mutations in the gene for hereditary diseases, (2) sporadic diseases with de novo mutations, (3) sporadic diseases with variations in disease-susceptible genes, and (4) sporadic diseases with other mechanisms. These different molecular bases are reviewed.

**APPARENTLY SPORADIC NEUROLOGICAL DISEASES WITH LOW-PENETRANT MUTATIONS IN THE GENE FOR HEREDITARY DISEASES**

There are numerous examples of low-penetrant mutations in apparently sporadic cases of neurological diseases. Sporadic cases of amyotrophic lateral sclerosis due to low-penetrant SOD1 mutations have been well characterized. In prion diseases, patients with V180I or M232R mutations in the prion protein (PRNP) gene rarely have a family history of prion diseases, indicating that these patients are usually diagnosed as having sporadic Creutzfeldt-Jakob disease.

**SPORADIC NEUROLOGICAL DISEASES WITH DE NOVO MUTATIONS**

Alternating hemiplegia of childhood is a rare neurological disorder characterized by early-onset episodes of hemiplegia, dystonia, various paroxysmal symptoms, and developmental impairments. Almost all cases are sporadic, but the concordance of alternating hemiplegia of childhood in monozygotic twins and the dominant transmission in a family with a milder phenotype have been reported. With this background information, Rosenwich et al conducted whole-exome sequencing of 3 proband-parent trios to identify a disease-associated gene and then examined whether mutations in the gene were also present in the remaining patients and their healthy parents. Whole-exome sequencing indeed showed 3 heterozygous de novo missense mutations. Similar approaches have been used for a number

---

**Figure 2.** Diagram showing the paradigm shift (ie, the explosive growth in genome science and medical genomics). Over the past decade, genome-wide association studies (GWASs) using common single-nucleotide polymorphisms (SNPs) have been conducted to identify genomic variations in sporadic neurological diseases. The theoretical framework of GWASs is the common disease–common variants hypothesis. Although GWASs have successfully revealed numerous susceptibility genes for common diseases such as diabetes mellitus, as well as neurodegenerative diseases, the odds ratios associated with these risk alleles are generally low and account for only a small proportion of estimated heritability. The availability of high-throughput genome sequencing technologies will enable us to identify all the genomic variants, and eventually those of disease-relevant alleles based on the common disease–multiple rare variants hypothesis.
of diseases, including severe epileptic encephalopathy, autism, and schizophrenia. The rationale for these approaches is based on the hypothesis that patients with severe phenotypes associated with reduced reproductive fitness may harbor de novo mutations.

Twin studies in which differences in the phenotypes of monozygotic and dizygotic twins were compared have long been conducted to delineate the involvement of genetic factors. Therefore, the comparison of whole-genome sequences of discordant monozygotic twins is expected to accelerate the discovery of genomic variations responsible for the disease phenotypes.

Over the past decade, genome-wide association studies (GWASs) using common SNPs have been conducted to identify genomic variations associated with sporadic neurological diseases. The theoretical framework of GWASs is the “common disease–common variants” hypothesis, in which common diseases are attributable in part to allelic variants present in more than 5% of the population. Although GWASs have successfully revealed numerous susceptibility genes for common diseases such as diabetes mellitus, as well as neurodegenerative diseases, the odds ratios associated with these risk alleles are generally low and account for only a small proportion of estimated heritability.

In GWASs, the general finding that the odds ratios associated with risk alleles identified for disease susceptibility are low indicates that GWASs based on the common disease–common variants hypothesis are not effective in identifying genetic risks with large effect sizes. The current experience with GWASs strongly suggests that rarer variants that are difficult to detect by GWASs may account for the “missing” heritability. Such rare variants may have large effect sizes as genetic risk factors for diseases. Thus, the paradigm should be shifted from the “common disease–common variants” hypothesis to the “common disease–multiple rare variants” hypothesis to identify disease-relevant alleles with large effect sizes (Figure 3).

An excellent example of rare variants with substantially large effect sizes is the recent discovery of the glucocerebrosidase (GBA) gene as a robust genetic risk factor for PD. A population-based study coupled with genealogy information demonstrated that the estimated risk ratio for PD for siblings of patients with PD was significantly high, indicating that genetic factors substantially contribute to the development of sporadic PD. Recent clinical observations have suggested the association of sporadic PD with heterozygous mutations in the GBA gene encoding the enzyme that is deficient in patients with Gaucher disease, an autosomal recessive lysosomal storage disease. Furthermore, the comorbidity of PD and Gaucher disease was previously described. We conducted an extensive resequencing analysis of GBA in patients with PD and controls, and we found that GBA variants that are pathogenic for Gaucher disease confer a robust susceptibility to sporadic PD and even account for the familial clustering of PD. The combined carrier frequency of the “pathogenic variants” was as high as 9.4% in patients with PD and significantly higher than that in controls (0.37%), with a markedly high odds ratio of 28.0 (95% CI, 7.3–238.3) for patients with PD compared with controls.

Figure 3. Diagram showing the road map to the identification of disease-relevant variations. Shifting the paradigm from the common disease–common variants hypothesis to the common disease–multiple rare variants hypothesis will lead to the elucidation of the molecular bases of sporadic neurological diseases. Relatively rare sporadic neurological diseases will be good candidates for identifying disease-relevant alleles with large effect sizes because, depending on the effect sizes, the sample sizes can be small. GWAS indicates genome-wide association studies; SNPs, single-nucleotide polymorphisms.
We can draw the following conclusions from the discovery of the major disease-susceptibility gene (GBA) with a large effect size: (1) a genetic factor with a large effect size has been discovered in sporadic PD; (2) in accordance with the large effect size, there is a tendency of familial clustering (multiplex families such as affected siblings); and (3) the disease-relevant allele could not be identified by GWASs using common SNPs and was identified only by nucleotide sequence analysis. These conclusions strongly encourage us to search for disease-susceptibility genes with large effect sizes based on the common disease–multiple rare variants hypothesis. Although the majority of rare missense variants have been suggested to be functionally deleterious in humans, it remains controversial whether a comparison of allele frequencies of rare variants (in particular, missense variants) is a sufficient method for identifying variants associated with diseases. Functional annotation of all the variants obtained by comprehensive genome sequencing will no doubt increase the robust power for detecting significant associations of variants with diseases.

**SPORADIC NEUROLOGICAL DISEASES WITH OTHER MECHANISMS**

Besides the mechanisms already mentioned, there may be others underlying sporadic neurological diseases. The involvement of somatic mutations occurring in certain cell lineages in sporadic neurological diseases is a potentially interesting mechanism. Such a mechanism in certain types of cancer is well established. The involvement of epigenetics in the development of sporadic neurodegenerative diseases is also a potentially attractive mechanism. Recently, there have been an increasing number of studies suggesting that “prion-like” processes (ie, the propagation of misfolded proteins leading to abnormal aggregation) may be involved in the pathogenesis of sporadic neurodegenerative diseases. In the field of autoimmune diseases such as multiple sclerosis, the involvement of genetic factors is well characterized. The application of massively parallel sequencing to extensively characterize T-cell receptor repertoires and immunoglobulin heavy chain genes along with sequence-based typing of HLAs will provide new insights into the molecular bases of autoimmune diseases.

As discussed in this review, the availability of robust technologies using NGSs will revolutionize our research paradigms for exploring the molecular bases of hereditary and sporadic neurological diseases. Furthermore, these technologies will soon be applied in clinical practice. It will be a new era of datacentric clinical practice. Are we prepared for this new era?


