Whole-genome association analyses have begun to yield confirmed findings for genetic risk variants for complex disease. As the first reports of its application to neurological disease are described, we review this progress, explain the principles of the analysis, and discuss what the future is likely to be in this exciting area.

Arch Neurol. 2008;65(3):319-321

The human genome draft sequence released in 2001 was a consensus sequence based on the stitching together of DNA sequences from clones derived from many individuals; at best, this corresponded to an imperfect sketch of the human sequence and certainly represented no one person. The immediate utility of the human draft DNA sequence was that it provided a map to allow scientists to localize genes that were mutated in mendelian disease. It did not directly help us to understand the more subtle differences between us, including predispositions to the many common diseases that afflict humans. These common diseases, which include most cases of neurological disease such as most amyotrophic lateral sclerosis, Parkinson disease, Alzheimer disease, stroke, brain tumors, many other cancers, most heart disease, and type 2 diabetes, have been believed to be predisposed to by many common variants across our genome. It has been believed that much human disease had its roots in individuals with unfortunate combinations of variants in different genes across their genome, perhaps with exposure to predisposing environmental factors and possibly a little bad luck, too.

Despite this pervasive belief, few of these common variants have been identified. Two approaches have been used during the last 15 years to find them: candidate gene association studies and affected family member (sibpair) linkage studies. While each has had limited successes, progress in general had been disappointing.

This last year, finally, the drought in genetic findings for complex diseases has ended and a deluge of clear disease associations has been reported. The reasons for this sudden change in fortune are both scientific and technological. The scientific change was the systematic identification of polymorphisms across the human genome. It has led to the discoveries that variability was not random in any population and that variability at one position could predict adjacent variability with reasonable accuracy (Figure). These realizations were systematized into a knowledge base of the human HapMap, which cataloged those single-nucleotide polymorphisms that could be used to capture the majority of human variability and was used to choose approximately 500,000 single-nucleotide polymorphisms whose genetic analysis could be used to assess about 93% of genetic variability. The technological advance was the development of 2 competing platforms that allow the assessment of these tagging single-nucleotide polymorphisms (http://www.affymetrix.com/index.axx and http://www.illumina.com/). The competition between the 2 platforms has driven the price down from approximately $1000 per individual to cover 50% of the genome in mid 2005 to approximately $250 to cover 95% of the genome in mid 2007.
as a candidate gene locus because it is the site of action of the insulin degrading enzyme (IDE) and the pancreatic β-cell transporter gene (SLC30A8) also met genomewide significance. Then, 3 other studies confirmed these findings and, individually and through pooling of their data, identified CDKAL1, CDKN2A/CDKN2B, FTO, KCNJ1, and IGF2BP2 as risk factor loci as well as confirmed PPARG (first identified as a candidate gene locus because it is the site of action of rosiglitazone maleate).

This story illustrates several points. First, large studies can find real associations (each study had on the order of 1,500 cases and controls). Second, replication leads to confidence. Third, pooling of data from the studies led to extra findings (3 studies pooled their data, and now, all of the groups who have published their findings are also pooling their data), leading fourth to the expectation that when the data represent approximately 15,000 cases and 15,000 controls, other findings will be made. Fifth, this approach can lead to directly “druggable” targets. As 2 examples, PPARG is the site of action of the major drug class for type 2 diabetes and zinc supplementation had already been considered as a therapy for diabetes.

To date, no confirmed findings to our knowledge have been reported for neurological diseases except the confirmation that the method picks up APOE in Alzheimer disease and the MAPT locus in progressive supranuclear palsy. However, there have been initial reports for Alzheimer disease, amyotrophic lateral sclerosis, Parkinson disease, and ischemic stroke. Most of these studies have publicly released their data, facilitating replication. No doubt, definitive findings will be made in the next period.

However, perhaps as exciting as this identification of single-nucleotide polymorphism associations has been the realization that this technology can pick up unexpected homozygosity and thus be used to clone recessive disease loci almost instantly. Finally, this technology also rapidly identifies the newly recognized large insertion, deletion, and inversion polymorphisms in the human genome that also have clear implications for the dissection of the pathogenesis of neurological disease.

Despite the phenomenal progress these findings represent, they are raising as many questions as they promise to answer. Many of the confirmed “hits” are not near genes: what do they represent? Distant control elements or RNA regulatory transcripts are possibilities. Most hits in genes do not alter the amino acid sequence, and this must represent variability that alters expression or splicing as we have recently shown is relevant for MAPT elements. Surprisingly, most of the new associations do not explain previous genetic linkage results, perhaps suggesting that much rare variability awaits to be found as is certainly true for cholesterol metabolism and may also be true in Alzheimer disease. Two other daunting tasks await geneticists: first, developing an understanding of gene–gene and gene–environment interactions, and second, developing the technologies (both technical and analytical) to generate and interpret the whole-genome medical sequencing that will be possible within the next 5 years. The progress this last year has been truly astounding. Much will be achieved in the next 2 years, but perhaps as important, this progress means that the even more difficult goals needed to fully understand disease pathogenesis now seem within reach. To attain these goals, we will need the resolve and self-confidence to collaborate and pool data, even with our competitors. We will need to recognize that these efforts require massive clinical and laboratory investments; thus, we must ensure that academic rewards and incentives are assured for all involved. Lastly, we will need the continued commitment of funding agencies because these experiments are both large scale and expensive; however, they promise to supply unequivocal answers to questions we have been asking for a long time and thus provide genuine value for money.

Figure. In this theoretical example, there are 4 single-nucleotide polymorphisms (SNPs) (in capital) that are close to each other and each of which has allele frequencies of approximately 50%. This might suggest that each of the 4 combinations would occur one-sixteenth of the time—approximately 6%—but in practice this is not what we see, and in fact 2 variants, 1 and 2, predominate. This means that if you genotype just 1 of the SNPs (for example, SNP2), you can guess with reasonable accuracy (but not certainly) what the other SNPs would be in that individual. Thus, SNP2 captures or tags most of the genetic diversity at this locus.

The result of this technological progress has been that increasing numbers of diseases are being analyzed by this whole-genome technology and are yielding confirmed risk factor loci. Early success in age-related macular degeneration, where the identified locus conferred a strong genetic risk, has been swiftly followed by efforts that have revealed and confirmed moderate and minor risk loci in type 2 diabetes, heart disease, atrial fibrillation, prostate cancer, and breast cancer. Progress is now so rapid that this list grows each week; the initial effect of these studies is that they immediately reveal “low-hanging fruit,” risk loci that have effects substantial enough to allow detection at first pass; as one commentary has recently put it, this has been likened to drinking from a fire hose. Increasing sample numbers will eventually allow the detection of very small effects and effects that are reliant on multigene or gene-environment interactions. Given this rapid progress, it is perhaps worth briefly reviewing the single example of type 2 diabetes as an illustration of the general principles involved and issues raised.

In 2006, using a linkage approach, the Decode group reported that the transcriptional gene TCF7L2 was a risk locus for this disorder. Sladek and colleagues confirmed this finding in a whole-genome study and additionally reported that the insulin degrading enzyme locus (IDE) and the pancreatic β-cell transporter gene (SLC30A8) also met genomewide significance. Then, 3 other studies confirmed these findings and, individually and through pooling of their data, identified CDKAL1, CDKN2A/CDKN2B, FTO, KCNJ1, and IGF2BP2 as risk factor loci as well as confirmed PPARG (first identified as a candidate gene locus because it is the site of action of rosiglitazone maleate).
REFERENCES


