Human Genome Sequence Variation and the Search for Genes Influencing Stroke

Jonathan Rosand, MD, MS; David Altshuler, MD, PhD

Background—Technological progress spurred by the Human Genome Project is accelerating the pace of genetic studies of common diseases, including stroke. Stroke clinicians will soon need to interpret increasingly complex genetic studies.

Summary of Review—Linkage analysis and epidemiological association are 2 fundamental methods of identifying gene variants affecting common diseases such as stroke. Combining these methods with advanced molecular genetic techniques, 3 recently published studies have made important contributions to the genetics of common vascular diseases: identification of the location of a gene for stroke on chromosome 5q12 and identification of gene variants that may increase risk of myocardial infarction. Driven by genomic technology, future studies will be increasingly comprehensive and systematic in their assessment of the contribution of genetics to the clinical course of stroke. The scale and complexity of such studies will require large-scale collaboration among stroke physicians, geneticists, and biostatisticians.

Conclusions—Rapid improvements in technology and study design are likely to elucidate the role of inherited genetic variation in complex diseases such as stroke. Understanding the methods of population-based genetic investigation and the patterns of human genome variation will enable stroke physicians to follow these future developments. **(*Stroke*. 2003;34:2512-2517.)**

Key Words: genetics ■ human genome project ■ stroke

Announcements of genetic discoveries are, for the first time, reporting results directly relevant to the care of patients commonly encountered by stroke physicians. Advancing technology is enabling the simultaneous measurement of more and more genetic information in larger and larger groups of patients. As a result, it is increasingly practical to study the role of genetics in complex human diseases such as stroke. Last year, deCODE genetics of Iceland published data indicating the chromosomal location of a gene influencing common forms of stroke.[1] More recently, 2 different Japanese teams used large-scale association screening to identify particular gene variants that might affect susceptibility to myocardial infarction (MI).[2,3] These studies break new ground, applying sophisticated technology to investigate the role of genes in the diseases we care for in the clinic and hospital every day.

What then does a stroke physician need to understand to follow this new generation of genetic investigations, evaluate their claims, and ultimately integrate genetic results into research and clinical practice? The goal of this review is to describe methods for population-based genetic investigation of common diseases, summarize current knowledge of human genome variation, discuss their recent application to the study of cerebrovascular and cardiovascular disease, and describe ways in which population-based genetic research in stroke is likely to evolve.

See Editorial Comment, page 2516

Linkage Analysis

Linkage analysis is used to identify the chromosomal location of gene variants influencing a disease and has successfully identified the locations of hundreds of genes for rare, monogenic disorders. Although linkage analysis generally requires families in which >1 individual is affected with the disease of interest, families can be either in the form of extended pedigrees or, more simply, in the form of pairs of siblings or other relatives of which both members or 1 member has the disease. (The "affected sib pair" design is the basis for the ongoing National Institutes of Health–funded Siblings With Ischemic Stroke Study [SWISS].[4]) In samples from these families, an evenly spaced map of DNA sequence variants (known as markers) is used to trace the inheritance of each copy of each chromosome. Chromosomal segments that do not influence disease segregate randomly according to Mendel's laws: 50:50 assortment of each copy to all offspring, regardless of disease. A DNA segment that influences dis-

Received March 3, 2003; final revision received May 5, 2003; accepted June 20, 2003.

From the Stroke and Neurocritical Care Units (J.R.) and Department of Molecular Biology and Diabetes Unit (D.A.), Massachusetts General Hospital, and Program in Medical and Population Genetics, Whitehead Institute/Massachusetts Institute of Technology Center for Genome Research (J.R., D.A.), Boston, Mass.

Correspondence to Jonathan Rosand, MD, Stroke and Neurocritical Care Units, Massachusetts General Hospital, VBK-811, 32 Fruit St, Boston, MA 02114. E-mail jrosand@partners.org

Stroke **is available at http://www.strokeaha.org** **DOI: 10.1161/01.STR.0000091844.02111.07**

ease, however, will not be inherited at random. Instead, the particular copy that, in each family, carries the disease-causing mutation will be shared among affected family members more often than would be predicted by chance. The likelihood that a particular genome segment is linked to disease is quantified with a LOD score (log-of-odds ratio). A LOD score ≥3.6 is generally considered the criterion for concluding that linkage exists between a genome segment and disease. Although linkage analysis can locate a segment of the genome that influences disease, the implicated region is typically quite large: millions to tens of millions of letters of DNA, spanning dozens to hundreds of genes.

Linkage analysis is the starting point of choice for rare, monogenic mendelian disorders. Its track record for complex, polygenic traits, however, has been much less successful. There are a number of fundamental challenges to the application of linkage analysis to complex traits. First, assembling multiplex families as well as pairs of affected siblings is difficult when the disease has an advanced age of onset and high mortality. (The advent of the Health Insurance Portability and Accountability Act [HIPAA] in the United States is certain to make this even harder.) Second, in the case of diseases such as stroke, power may be diminished because the clinical disease classification (eg, Trial of Org 10172 in Acute Stroke Treatment [TOAST] stroke subtypes[5]) bears an imperfect relationship to the underlying biological (inherited) disease mechanisms. Third, linkage is an indirect statistical test, relying on distortions of mendelian inheritance ratios to infer the nearby location of a disease-causing mutation. When the genetic effect is very large (as in mendelian disorders), this indirect signal is sufficient. For assessment of modest effects and for variants that are common in the population, the power of linkage may simply be inadequate.[6] Fourth and finally, even where linkage is successful, it only identifies a region of interest rather than a causal mutation that might reveal biological insight or have clinical utility. For monogenic traits, "fine-mapping" and mutation hunting have been adequate to take the next step and identify most genes of interest. For complex traits, in contrast, only in the last 2 years have any investigators been able to go from linked region to gene.[7]

The study by deCODE Genetics used linkage analysis to search for a stroke susceptibility gene in 179 families containing 476 patients with ischemic stroke of any subtype or intracerebral hemorrhage.[1] Despite pooling distinct clinical entities, they developed significant evidence that 1 or more genes influencing these common forms of stroke exist within a region of chromosome 5q12. The region was named STRK1 by the investigators, although no particular culprit gene or mutation has yet been published. Until such a gene or gene variant is identified, it is not yet possible to glean pathophysiological or diagnostic information from this result. Moreover, on the basis of the magnitude of the effect observed (and because stroke is not a monogenic trait), the gene that may reside in the STRK1 region is unlikely to explain more than a modest fraction of risk of stroke in the Icelandic or other population. Nevertheless, narrowing the search for a stroke gene to a roughly 16 million base-pair region of DNA (a reduction of 200-fold compared with the 3 billion bases of DNA in the human genome) is a significant and promising advance. The next step in genetic analysis, now that a linked DNA region has been found, is to search for associations between particular DNA variants and risk of disease.

Association Studies

Association studies examine the frequency of specific DNA variants (alleles) in groups of unrelated individuals with disease and unaffected controls. Rather than tracking coinheritance of a chromosomal segment among affected individuals in a family (as in linkage analysis), association studies consider the inheritance of genetic variants within the population at large. Their main advantages are as follows: (1) they have greater statistical power than linkage analysis,[6] and (2) they do not require family-based collections. Demonstration of association, however, is not by itself sufficient evidence for a causative role of the gene variant studied. If the pathogenic polymorphism lies very close to the polymorphism studied, then it is possible that the studied polymorphism may be associated with disease simply because it is in "linkage disequilibrium" with the causative gene variant and is therefore inherited with it. Only studies in the laboratory that confirm altered function of the identified gene can ultimately confirm a role in disease.

Association studies use case-control or family-based designs to demonstrate association in the population between possession of a particular allele and disease (eg, apolipoprotein E $\epsilon 4$ and increased risk of Alzheimer disease). Case-control designs are no different in conception from the case-control methods that have been well developed for use in epidemiological studies. Cases and controls are enrolled from the same source population but are unrelated. Association designs that use family members as controls, such as the transmission disequilibrium test, have been developed to account for the potential confounding effect of population stratification. Population stratification occurs when cases and controls are unintentionally included at different ratios from ≥2 subgroups that have different ethnic or genetic backgrounds. In this case, a polymorphism that happens to be associated with ethnicity/genetic heritage (rather than disease) might appear to associate with disease. Recently described genetic methods promise to control for population stratification and have been applied in some studies, including that of Ozaki et al,[3] discussed below. These methods may render study designs like the transmission disequilibrium test unnecessary.

The main limitation of association studies is that they require a priori knowledge of putative mutations. In other words, investigators must identify "candidate genes" of interest on the basis of inherently incomplete knowledge of the biology of disease. Until recently, this limitation, combined with the restricted number of known genes, rendered association studies incomplete, arbitrary, and often irreproducible. Fortunately, however, the situation is changing. Advances in technology and our understanding of human genetic variation are allowing broader and more systematic surveys of possible genetic contributors to disease.

Genetic analyses from populations across the globe have revealed that the human genome contains very limited vari-

ation: any 2 copies drawn at random from the world's populations will differ at only 1/1250 bases examined.[8] In regions of the genome that code for proteins, variation is even more restricted, totaling only 1/2000 bases compared.[9–11] Moreover, the vast majority of differences between any 2 copies of the human genome are due to variants that are common in the population, that is, approximately 90% of the genetic variation in each of us is due to variants that are also found at a frequency >1% in the general population.[8,12,13] Recent efforts within the genomics community have sought to catalog common variants in the human genome, and it is now estimated that perhaps 4 million of the estimated 10 million common human sequence variants are now present in public and for-profit databases.[8,14]

The wide distribution of limited genetic variation is a byproduct of human population history: until a very short time ago (10 000 to 40 000 years), the human population was small (perhaps 10 000 individuals) and localized entirely within Africa. We are all descendents of that founder population. The bulk of human genetic variation is due to a modest number of common variants that were inherited from this population and are present all over the world. The remaining millions of rare variants have occurred more recently and are each found in a small number of people.

Evolutionary theory suggests that the common variants currently being cataloged around the world may be the DNA variants that contribute to the familial risk of complex diseases. In general, because mutations that are now common in the population are typically quite old (having occurred tens or hundreds of thousands of years ago), they are likely evolutionarily neutral or beneficial. Gene variants that are deleterious, in contrast, are likely to stay rare or to be lost as a result of natural selection.

The most plausible evolutionary impact of a particular disease may therefore suggest which class of variants, those that are common or those that are rare, are the best candidates for study. Most monogenic disorders are typically severe and manifest before reproduction. Mutations causing these diseases thus tend to be rare. Although the expected frequency of gene variants that influence common and late onset diseases remains controversial in the genetics community, one simple model is that the mutations causing common, late-onset diseases are likely to be evolutionarily neutral (explaining the relatively high frequency of the disease), and as a result they will often themselves be common in the population. While the relative contribution of common gene variants will probably be different for each disease, past natural selection is certain to have played a major determining role in the frequency and number of disease-causing mutations for all of them.

Application to Disease

Progress in characterizing and cataloging human genetic variation will allow increasingly comprehensive searches for disease-associated gene variants in stroke and other diseases. No large-scale association studies in cerebrovascular disease have yet been published, but 2 such recent studies in MI were important demonstrations of progress toward broader surveys of genetic variation for its relationship to disease.[2,3]

The first study, by Yamada et al,[2] is remarkable for its size, but its limitations are particularly instructive. The study examined polymorphisms in 71 genes in >5000 patients divided among MI cases and controls. Genes and variants were each selected on the basis of a hypothesis about biological processes contributing to coronary disease. Their findings, that variants in connexin 37, plasminogen-activator inhibitor type 1, and stromelysin-1 were associated with MI ($P<0.001$), is interesting and hypothesis-generating but must be treated with caution. Men and women were analyzed separately, with different results obtained in each sex. (The finding of sex-specific effects is certainly possible but introduces an opportunity to find spurious results in the data.) In addition, no evidence is presented about the functional consequence of the mutations studied. Finally, the level of statistical significance is only modest in light of the large number of variants in the genome and the low prior probability that any play a role in disease (see below).

The second study, by Ozaki et al,[3] was much more broad and systematic in approach and, in addition, achieved a stronger level of statistical significance. The authors had previously sequenced >13 000 human genes to identify >90 000 polymorphisms. Of these, 65 671 were tested for an association with MI in a small sample of 94 cases and 658 controls. Results with a $P<0.01$ were then tested in a much larger sample of >2000 cases and controls, and 2 variants in the lymphotoxin-α gene were shown to be associated with MI ($P<0.000003$).[3] The authors further demonstrated that these lymphotoxin-α variants altered function of the encoded protein, providing crucial support for their effect on the disease process. Despite its strengths, this study, like that of Yamada et al,[2] still requires replication.

Interpreting Association Studies

The validity of these straightforward association studies depends on the selection of appropriate controls. Misclassification of controls risks a false-negative result, particularly when the disease under study is itself relatively common in the population. For diseases such as stroke or MI, which tend to have an advanced age of onset, it is possible that controls may indeed be presymptomatic rather than free of disease. Control patients may even have asymptomatic disease (eg, small cerebral infarctions visible on CT scan only). Phenotypic characterization of controls as well as cases is therefore crucial and may require expensive diagnostic procedures such as brain imaging for controls as well as cases. Similarly, proper categorization of family members may also require such procedures. Unfortunately, practical considerations such as the cost of such undertakings as well as the restrictions on the flow of healthcare information recently mandated by HIPAA may present formidable obstacles to such extensive approaches.

A central challenge is to determine whether an observed difference in polymorphism frequency between cases and controls reflects true association rather than a chance event. This is particularly critical as we move from hypothesis-driven research to more unbiased genome-wide studies like that of Ozaki et al.[3] The genome is very large, and the number of true genetic risk factors is likely small in comparison.

Thus, the probability that any individual polymorphism tested actually influences disease is correspondingly low. Probability values of 0.05 are therefore inappropriate for such hypothesis-generating studies, and much more rigorous criteria need be applied. For example, we might hypothesize that there are 10 000 000 single-nucleotide polymorphisms in the genome, and 20 of these might truly influence a disease in a manner that could be detected in a given study design. With a probability value of 0.05, there would be 500 000 false-positive results and 20 real positive results. With a probability value of 10^{-6}, in contrast, there would be only 10 false-positives along with the 20 true positives. While there is no consensus on what represents a sufficiently conservative probability value for studies testing multiple hypotheses or low prior probabilities, investigators must consider all available evidence to decide whether the statistical results are likely to represent a true effect rather than statistical noise.[15]

Replication in multiple independent studies is at present the most reliable method of identifying a true relationship between polymorphism and disease. Unfortunately, however, few previously published genetic associations meet this criterion. When recently reviewed, the rate of replication of candidate gene association studies was strikingly low: of 166 associations studied >3 times, only 6 were consistently replicated.[16] There are 2 likely explanations for this lack of replication. Not only did these studies lack sufficient power (ie, included too few patients) to identify true positive findings among the great sea of possibilities, but they also applied statistical thresholds that were insufficiently rigorous to exclude false-positive findings.[17] Although the process is slow, the most suggestive associations are ultimately replicated many times, and eventually robust and reproducible findings emerge. Assembling large patient collections with adequate power to distinguish the few true associations from the sea of statistical fluctuations is an attractive alternative to awaiting replication and an approach the stroke community must consider.

Implications for Stroke Genetics

What is the likely application of this research to stroke? The deCODE report, while encouraging in its single success, is also sobering. Despite the large size of the study, only a single locus was found that is likely to explain only a fraction of disease. Better diagnostic categorization, much larger sample sizes, or more powerful methods will be required to uncover the other genetic risk factors for stroke.

To pursue statistically more powerful association studies, future investigations will have to account for the fact that the role of any given gene variant in the course of stroke is likely to be small. Several strategies may improve the probability of detecting genetic effects. Selection of patients on the basis of younger age of onset may help to identify those in whom genetic effects are stronger,[18] although the genes identified, as in the case of *BRCA-1* and *BRCA-2* and breast cancer, may demonstrate limited effect outside of early-onset disease. The application of uniform and biologically meaningful ways of categorizing stroke subtypes should facilitate the detection of subtype-specific genetic effects.[19] To this end, sophisticated observations from clinician researchers will be essential for characterizing stroke phenotype, including presumed biological mechanism, severity, response to treatment, and functional recovery, to name just a few characteristics that may be powerfully influenced by genetic factors. Indeed, identification of risk genes may, in fact, improve diagnostic categorization when patients with a common phenotype, such as cardioembolic stroke, are divided according to whether they possess a particular risk allele, propelling an iterative process of genetic investigation. Finally, the use of so-called endophenotypes or intermediate phenotypes, such as carotid atherosclerosis, which link genotype and the more complex phenotype of clinical interest, may result in more straightforward genetic analysis.[20] Presumably because simpler genetic effects underlie their development, endophenotypes may offer a shortcut to the discovery of important disease genes.

Which genetic variants should we systematically examine? One answer relates to the evolutionary effects, and hence allele frequencies, of the variants that might contribute to stroke.[6,21] On the one hand, stroke nearly always develops after reproduction has occurred. Its evolutionary effect may not be deleterious. Common gene variants may therefore offer one route to finding genes with a role in stroke. In contrast, it is possible that the cardiovascular and hemostatic factors that contribute to stroke may have other pleiotropic effects on survival and thus may be deleterious in a different context. If this were the case, then the variants underlying disease would tend to be rare, making studies of common variants less valuable and studies of rare variants more important. Only systematic studies of common variants and their relationship to stroke, using methods like that of Ozaki et al,[3] will help to resolve these possibilities.

Collaborating for the Future

While rapidly advancing technology will facilitate progressively broader assessments of multiple genes for contribution to disease,[22] successful identification of genes influencing stroke will depend on the stroke community's collaboration and improved methods of clinical characterization. It is likely that large sample sizes—the sort achieved only by collaboration and sharing of clinical material—will be required to obtain the needed statistical power. (Recall that Yamada[2] and Ozaki[3] and their colleagues included thousands of patients in their studies of MI, a more common and homogeneous disorder than ischemic stroke.) To accumulate the necessary sample sizes, cooperation among many centers will be essential. This will undoubtedly necessitate working together to define phenotypes, stroke risk factors, and outcome in a unified manner. Funding agencies will have to be convinced of the value of assembling these cohorts, even for exploratory genetic association studies. Of course, it will be vital not only to assemble the initial cohorts to test exploratory hypotheses but also to assemble future cohorts to allow independent confirmation of preliminary findings.

The formation of large research teams will have important consequences for the ways in which academic centers assign credit for academic promotion. Genetic research groups will require that specialists in vascular neurology work hand in hand (and share credit) with geneticists, statisticians, specialists in bioinformatics, and experts in high-throughput

genomic methods. Long author lists will contain the names of multiple contributors, many of whom will have made indispensable contributions that deserve legitimate recognition.

Conclusion

Despite years of progress in basic and clinical investigation, the pathogenesis of the most common stroke subtypes remains poorly understood. Genetics can offer powerful clues, and the technology to investigate these clues is developing quickly. Systematic association studies testing the range of common genome variants may identify the genes affecting susceptibility to stroke as well as its clinical course. Successful identification of stroke genes will require the collaboration of large numbers of neurologists and other clinicians who can use their expertise in clinical characterization to identify the clinical subtypes and aspects of disease course most likely to be affected by variation in the human genome.

Acknowledgments

This work was supported by the American Academy of Neurology Education and Research Foundation, National Stroke Association, and National Institute of Neurological Disorders and Stroke (National Institutes of Health grant 1 K23 NS42695-01). We thank Jose Florez, MD, PhD, Steven M. Greenberg, MD, PhD, and Christopher Newton-Cheh, MD, for helpful discussions and critical review of the manuscript.

References

1. Gretarsdottir S, Sveinbjornsdottir S, Jonsson HH, Jakobsson F, Einarsdottir E, Agnarsson U, Shkolny D, Einarsson G, Gudjonsdottir HM, Valdimarsson EM, et al. Localization of a susceptibility gene for common forms of stroke to 5q12. *Am J Hum Genet*. 2002;70:593–603.
2. Yamada Y, Izawa H, Ichihara S, Takatsu F, Ishihara H, Hirayama H, Sone T, Tanaka M, Yokota M. Prediction of the risk of myocardial infarction from polymorphisms in candidate genes. *N Engl J Med*. 2002; 347:1916–1923.
3. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, Sato H, Hori M, Nakamura Y, Tanaka T. Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat Genet*. 2002;32:650–654.
4. Meschia JF, Brown RD Jr, Brott TG, Chukwudelunzu FE, Hardy J, Rich SS. The Siblings With Ischemic Stroke Study (SWISS) protocol. *BMC Med Genet*. 2002;3:1.
5. Adams HP Jr, Bendixen BH, Kappelle LJ, Biller J, Love BB, Gordon DL, Marsh EE III. Classification of subtype of acute ischemic stroke: definitions for use in a multicenter clinical trial: TOAST: Trial of Org 10172 in Acute Stroke Treatment. *Stroke*. 1993;24:35–41.
6. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273:1516–1517.
7. Hugot JP, Chamaillard M, Zouali H, Lesage S, Cezard JP, Belaiche J, Almer S, Tysk C, O'Morain CA, Gassull M, et al. Association of nod2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*. 2001;411:599–603.
8. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 2001;409:928–933.
9. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*. 1999;22:231–238.
10. Li WH, Sadler LA. Low nucleotide diversity in man. *Genetics*. 1991; 129:513–523.
11. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat Genet*. 1999;22: 239–247.
12. Marth G, Yeh R, Minton M, Donaldson R, Li Q, Duan S, Davenport R, Miller RD, Kwok PY. Single-nucleotide polymorphisms in the public domain: how useful are they? *Nat Genet*. 2001;27:371–372.
13. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. The structure of haplotype blocks in the human genome. *Science*. 2002;296:2225–2229.
14. Reich DE, Gabriel SB, Altshuler D. Quality and completeness of SNP databases. *Nat Genet*. 2003;33:457–458.
15. Dahlman I, Eaves IA, Kosoy R, Morrison VA, Heward J, Gough SC, Allahabadia A, Franklyn JA, Tuomilehto J, Tuomilehto-Wolf E, et al. Parameters for reliable results in genetic association studies in common disease. *Nat Genet*. 2002;30:149–150.
16. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med*. 2002;4:45–61.
17. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet*. 2003;33: 177–182.
18. Kittner SJ. Stroke in the young: coming of age. *Neurology*. 2002;59:6–7.
19. Meschia JF. Addressing the heterogeneity of the ischemic stroke phenotype in human genetics research. *Stroke*. 2002;33:2770–2774.
20. Gottesman II, Gould TD. The endophenotype concept in psychiatry: etymology and strategic intentions. *Am J Psychiatry*. 2003;160:636–645.
21. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001;69:124–137.
22. Tsuchihashi Z, Dracopoli NC. Progress in high throughput SNP genotyping methods. *Pharmacogenomics J*. 2002;2:103–110.

Editorial Comment

The Pendulum's Swing: The Way Forward in the Genetics of Stroke

In a field that has been paved with frustration because of inconsistent reports, the review article by Rosand and Altshuler[1] comes at the right time. It is refreshing to see that the authors rightly adopt an optimistic view on the prospects of molecular genetic studies in unraveling the molecular and genetic architectures of stroke.

In a nutshell, the goal of molecular genetic investigations is to identify genetic mutations that confer an individual's genetic susceptibility to the disorder. The identification of causative mutations forms the basis for diagnostic and prognostic tests. It also allows us to comprehend the molecular etiology and pathophysiology of the disorder, which in turn serves as a springboard for the development of therapeutic modalities that are tailored to the yet-to-be-discovered molecular abnormalities. So much for the easy part. Stroke indeed, as all other complex disorders, is underscored by the combined effects of several to many genes with reduced penetrance. Furthermore, it is likely that different sets of deleterious genes contribute to the disease in different populations or families. How we should go about identifying susceptibility genes remains a burning issue.

Twenty years ago, the general opinion was that, if researchers wanted to characterize genetic effects of diseases (of monogenic types), they had to strictly adhere to linkage analysis procedures. One of the revolutions of molecular geneticists in the 1980s was to show that modern tools of recombinant DNA technologies could be applied to the unraveling of complex (then called *multifactorial*) disorders, whereby polygenic factors interplay with environmental and epigenetic factors. Those common disorders, which are chronic and degenerative in nature, include atherosclerotic diseases, hypertension, diabetes, allergies, cancers, Alzheimer disease, and stroke, among others. The higher degree of complexity detracted many classical geneticists, and association studies in particular were largely frowned upon. In the mid-1990s, benefits and shortcomings of different strategies were reasonably assessed.[2,3] The result was an explosion of association studies to the extent that the pendulum seemed to have swung from "purely linkage" to "exclusively association" lately.

In fact, both general strategies, together with variations on their themes, are complementary, and investigators should be very much aware of advantages but also of limitations and pitfalls of each of them. Rosand and Altshuler's balanced view on the topic is even more so remarkable that, compared with other complex clinical entities, stroke is clearly more amenable to case-control types of investigations because of both late age of onset and associated mortality. And although both linkage[4] and case-control studies[5] have so far yielded positive results, they still fall short of giving definite answers.[1] Clearly, there is a need for better diagnostic categorization, larger samples sizes, and more powerful methods.[1] The review gives practical ways of avoiding some of the pitfalls of association study designs.

For example, usually accepted P values of 0.05 are clearly no longer good enough, and sample sizes should be sufficiently high to reach statistical significance down to an order of 10^{-6} instead. Another key element is the replication of data and, even better, to carry out meta-analyses. Adequate samples sizes can better be achieved through multicenter study designs, with obvious implications on how to share and assign due credit to all collaborators. Then, while the selection of patients with well-defined clinical criteria is a critical issue, the major difficulty resides in the recruitment of controls. These are usually age- and sex-matched individuals who are at best disease-free at the time of sampling, although they may very well be presymptomatic already. Collectively, they therefore should be more appropriately referred to as comparison (rather than control) groups. While focusing on end-point clinical phenotypes (such as stroke), probing for associations with endo- (intermediate) phenotypes will yield critical information to explain underlying pathophysiological mechanisms. Two other topics would also deserve special attention: (1) Besides drastically increasing sample sizes, another way to improve the power of analyses is to construct haplotypes (combinations of markers on the same chromosomal region) as opposed to using single markers in order to define alleles on which to then look for causative variants.[6] At several loci indeed, linkage disequilibrium has been shown to be lost after 3.5 kilobase pairs. (2) The identification of culprit genes in chronic conditions relies on 2 basic assumptions, those of clinical and genetic homogeneities. While the idea of defining strict yardsticks for inclusion of cases and controls is well accepted, genetic make-ups of sample populations under investigation are rarely questioned. Yet genetic isolates or specific ethnicities would clearly be more appropriate at first than populations of mixed, recent genetic origins, to tease out major genetic effects.

Clearly the time is ripe to establish proper sets of guidelines that will arm investigators with convincing criteria to identify disease-susceptibility genes. We have now learned enough lessons, and there has to be law and order to prevent wastage of time and resources.

The review even touches upon evolutionary considerations. The authors give preference to the model stating that " . . . mutations causing common, late-onset diseases are likely to be evolutionarily neutral . . . ," although the usual view is that we have inherited those genes that allowed our ancestors to survive under harsh conditions. Examples include the "thrifty gene" hypothesis for diabetes, and the fight or flight response leading to increased blood pressure, and through the same molecular pathways, to essential hypertension. Similarly, the mechanisms developed throughout evolution for storing and retaining as much energy as possible from very limited food supplies have turned against individuals of affluent societies, in whom relatively recently improved environmental conditions lead to early development of atherosclerosis. In the case of stroke, which is the fourth leading cause of death in the world (and the third one in developed countries), Rosand and Altshuler[1] reconcile the 2 views in arguing that the molecular pathways leading to cerebrovascular accidents may have exerted pleiotropic effects on survival. Selective advantages could find their roots in mechanisms similar to those leading to atherosclerosis.

No review could possibly cover all important aspects of the field of genetics of complex disorders. Bearing this reality in mind, Rosand and Altshuler have done an excellent job at producing a balanced, synthetic overview of genetics and stroke. They are encouraging readers to raise their awareness of underlying concepts and methods, thereby alerting them that efficiently generated results are quickly crossing the doors of research laboratories to enter clinical practice.

Muhammad Ali Bangash, MD, *Guest Editor*
Philippe M. Frossard, PhD, DSc, *Guest Editor*
Department of Biological and Biomedical Sciences
Aga Khan University Medical College
Karachi, Pakistan

References

1. Rosand J, Altshuler D. Human genome sequence variation and the search for genes influencing stroke. *Stroke*. 2003;34:2512–2517.
2. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*. 1994; 265:2037–2048.
3. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273:1516–1517.
4. Gretarsdottir S, Sveinbjornsdottir S, Jonsson HH, Jakobsson F, Einarsdottir E, Agnarsson U, Shkolny D, Einarsson G, Gudjonsdottir HM, Valdimarsson EM, et al. Localization of a susceptibility gene for common forms of stroke to 5q12. *Am J Hum Genet*. 2002;70:593–603.
5. Hassan A, Markus HS. Genetics and ischaemic stroke. *Brain*. 2000;123: 1784–1812.
6. Akey J, Jin L, Xiong M. Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet*. 2001;9:291–300.

Human Genome Sequence Variation and the Search for Genes Influencing Stroke

Jonathan Rosand and David Altshuler

Stroke. 2003;34:2512-2516; originally published online September 18, 2003;
doi: 10.1161/01.STR.0000091844.02111.07
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231

Print ISSN: 0039-2499. Online ISSN: 1524-4628