Where Do We Go for Atherothrombotic Disease Genetics?

Stefan-Martin Brand-Herrmann, MD, PhD

See related articles, pages 1109–1120.

Cardiovascular disease (CVD), including coronary artery disease (CAD) with its fatal clinical outcome myocardial infarction (MI) and stroke, remain the leading cause of deaths in western and nonwestern countries. In the US, CVD claims more lives than cancer, chronic lower respiratory tract diseases and diabetes mellitus combined, accounting for ≈36% of all deaths in 2004 and incurring an estimated direct and indirect cost of ≈432 billion dollars in 2007.

Stroke and MI are multifactorial traits with complex pathophysiology. Besides other established factors, family history of disease strongly and independently determines CVD risk. There is a nearly 5-fold difference in stroke risk between monozygotic and dizygotic twins, and with respect to race, stroke incidence is approximately twice as high in blacks compared to whites. Marenberg et al, in their study on ≈20 000 Swedish twins, demonstrated that the relative risk to die from MI, when the other twin already died from this disease, was twice as large for monozygotic compared to dizygotic twins.

Celera Genomics and the International Human Genome Project consortium, by whole genome shotgun and clone-based sequencing strategies, respectively, proposed 2 versions of the human genome—the diploid genome offering more genetic information than previously anticipated. The availability of almost the entire human genome sequence allows for the immediate access to specific target sequences across the human genome. The number of different technologies to explore genetic sequences on a “genome-wide” scale and high-throughput genotyping methods is steadily increasing. Genome-wide approaches are essentially based on the fact that alleles at nearby loci often show strong linkage disequilibrium (LD). Human recombination events occur ubiquitously in short genomic hotspots (1 to 2 kb) every 100 to 200 kb, preferentially outside genes, leading to a breakdown of allelic association. Only a limited number of markers within each region of strong allelic association is necessary to ‘tag’ nearby genetic variation (SNP tagging approach). To map the structure of allelic association across the human genome is one of the main goals of the International HapMap Project.

There are roughly 2 major approaches to study the contribution of genetics to complex diseases, depending on the availability of population structures, ie, unrelated individuals or families: (1) association studies and (2) linkage analyses. Both approaches can be subdivided into the (1) candidate gene approach (hypothesis-driven) and very recently (2) genome-wide approaches (nonhypothesis-driven). The traditional candidate gene approach is largely based on our actual knowledge of biological pathways and metabolic interactions that may play a role in the pathophysiology of, for example, atherosclerosis, with a more specific focus on its inflammatory component. All genes encoding products for vasoactivity, lipid metabolism, thrombosis, atherosclerosis, and others, represent potential candidates for both MI and stroke, ie, atherothrombotic disease. Other candidate genes—without prior hypotheses—can be identified via distinct gene/transcriptome expression techniques using microarrays or proteomics in appropriate cell types and tissues at different pathological stages (direct approach) or via marker studies based on genome-wide scans (linkage or association).

Because most studies identified only modest association of common genetic variants with complex traits, which was supported by a recent meta-analysis of 301 association studies, and most of these studies relied on the candidate gene approach, genome-wide approaches offer the opportunity to identify additional risk loci, even if this does not warrant the identification of the gene(s), which are deemed responsible for the significant genomic signals. If no biologically plausible or even no annotated genes can be identified in the regions of interest, and also exhaustive downstream resequencing of large regions adjacent to the candidate locus fail to identify the responsible gene, the interpretation of results remain obscure.

In this issue of Stroke, 2 groups address the concept of genetic contribution of candidate loci or genes to stroke very differently. Sherva et al, on the one hand, performed a genome-wide strategy using affected sib-pairs and different linkage models to detect regions that are potentially linked with atherothrombotic disease defined as the composite phenotypes of stroke or MI, based on classification via self-report. The approach was based on 9607 individuals in 226 black, 395 Hispanic, and 207 white families; 106 families had multiple affected individuals. In their analysis, 880 patients had stroke or MI, and 83 patients reported both outcomes. The authors observed significant linkage signals for MI or stroke in Hispanics on chromosomes 2q21 and 7q21.1, in blacks and Hispanics on 9q32, in blacks on 11p13, and finally in whites on 12q24.33. Interestingly, the investigators replicated former linkage results for chromosome 2q21.1 to 22, even if in distinct ethnic populations than
previously published by others; the linkage was significant in both blacks and Hispanics, but not in their white sample. If the results were correct, it would strengthen the concept of genetic associations across ethnic populations. Besides other hypotheses, the authors emphasized that the 2q21.1 to 22 region contains the thrombospondin type-1 domain-containing protein 7B precursor gene (THSD7B), which encodes a protein with unknown function. They further suggested similar actions as known for thrombospondin type-1 (TSP1), which indeed seems to be involved in CVD. However, THSD7B is a single-pass type I membrane protein of 1608 amino acids (179402 Da), and is merely characterized as possessing a TSP1 domain. At the genome level, THSD7B is related to THSD7A and the pseudogene LAT1-3TM; so overall it does not appear as obvious candidate for cardiovascular (patho)physiology.

With regard to the 3 different populations under study, many factors, including plasma cholesterol level, differed significantly even if relatively modestly, across populations. However, the most apparent difference was attributed to the prevalence of diabetes mellitus, which was 60.9%, 17.9% and 10.7%, in Hispanics, blacks and whites, respectively. This prevalence of diabetes mellitus, which was 60.9%, 17.9% and 10.7%, in Hispanics, blacks and whites, respectively. This difference may reside on distinct haplotypic background constellations, which may explain their inconsistent associations. This necessitates appropriate resequencing of that region with subsequent in-depth functional profiling of the suspected genetic variants. A molecular functionality of the above mentioned potential candidate variants has not yet been demonstrated.

With regard to the choice of candidate variants, authors used the following inclusion criteria: (1) location of the variant, (2) minor allele frequency >5%, and (3) significant evidence from previously published studies. The last criterion is perhaps the less disputable, given that previous results were correctly interpreted as plausible evidence of association, and not too many conflicting findings existed that are opposed to previous evidence. However, this strongly relates to the issue of replication (see below).

With respect to the location of the variant, if it is not merely related to LD patterns, it is very difficult to qualify a potential functionality of a genetic variant simply from its localization on the gene. A molecular biologically plausible functional variant may reside in any gene region; the most convincing criterion being its proven functionality as has recently been very instructively shown even for intronic and synonymous or “silent” variants, which do not affect the amino acid sequence of the protein. Indeed, synonymous variants have been shown to affect mRNA splicing or stability, or even translation efficiency giving rise to an altered protein folding. More complicating even, the overall “net” function of genetic variation may depend on the combination of variants as molecular haplotypes. As shown for the dopamine receptor D2 (DRD2), the effect of a synonymous variant on decreasing mRNA translation and stability, and on weakening the response to dopamine-induced upregulation of DRD2, could be amnullied by the presence of a second synonymous variant (combined clone, ie, on the same molecular haplotype), which had no functional impact on its own.

Even if a minor allele frequency of >5% is a relatively common and practicable way of choosing variants to be typed in a given population to provide a reasonable power, most genetic polymorphisms occur within a frequency range of 1%
to 10%.\textsuperscript{31} and very often functional missense mutations are found with a frequency \(\approx 1\%\). It is now widely considered that genetic variants with allele frequencies <5\% may significantly contribute to complex diseases\textsuperscript{32,33} and studies have been directed specifically toward the discovery of rare variants.\textsuperscript{34}

As clinical doctors, we are asked to be cautious in interpreting results relating relative risk and the estimate of effect sizes to phenotype expression: for example, the development of stroke or MI. Indeed, relative risk should not be confounded with absolute risk,\textsuperscript{35} and a clear distinction can prevent from misleading conclusions. Relative risk gives a prospective—and odds ratio a retrospective—estimate of, for example, the occurrence of a clinical event in these individuals compared to those without this factor. It is good practice to “search” for a given factor in a retrospective study and to confirm more specifically and with adequate power in a prospective study. Population attributable risk provides an estimate of how many individuals develop a phenotype specifically due to the presence of the factor. In that respect, population attributable risk largely relies on the frequency of the candidate factor in the population. In terms of genetics, the allele frequency would be the driving factor. Genotype-phenotype relationships should certainly be differently interpreted compared to environmental-exposure-phenotype relationships, ie, smoking and lung cancer risk, which is one of the most obvious examples. Another criterion may be the effect size of the observed association. A functionally relevant variant may impart a higher effect size than a nonfunctional variant in incomplete LD with the functional one.

As the authors also allude to in their report, replication is an “essential prerequisite of any candidate gene study of this nature.” Replication can be strictly defined as association with the same allele, the same trait under the same genetic nature.” Replication can be strictly defined as association with the same allele, the same trait under the same genetic model.\textsuperscript{36} Replication may also apply to the candidate loci as such in a broader genomic context or more specifically to the same candidate risk alleles in the sense of “best-associated” variant. The most important criterion related to this issue is the molecular biological functionality of the variant of interest. If it has been undoubtedly shown that a variant confers an in vivo/in vitro functionality, which complies with the phenotype of interest, future association results should be replicated with that specific variant,\textsuperscript{37,38,39} even if it did not appear as ‘best-associated’. When only a marker in LD with the potential functional variant has been identified, association results might vary from one ethnic population to another, depending on the strength of LD among these variants (see LTBR variants in this editorial). It is certainly still a matter of debate of how many times a genetic association should be replicated to finally being accepted as such. Or if a formerly classified molecular functional variant does not prove to be replicated across populations,\textsuperscript{40,41} which criteria allow us to disqualify the variant? From some of these examples, if it was not purely related to differences in population structure, we learned that sources of inconsistencies include the lack of accounting for context-dependencies or gene-environment interactions.\textsuperscript{42,43}

Cerebral infarction, the most common form of stroke, is a quite heterogeneous multifactorial disorder and can be classified into the following subtypes: (1) lacunar infarction due to arteriosclerosis of small arteries and without any finding in favor of an atherothrombotic or cardioembolic stroke; (2) atherothrombotic infarction due to atherosclerosis of the external and major intracranial arteries; (3) cardioembolic infarction due to a cardiac embolus; (4) arterial dissection with typical clinical-angiographic patterns of carotid or vertebrobasilar artery dissection; (5) other etiologies including polycythemia vera, cerebral arteritis, or thrombocytopenia; (6) undetermined cause when \(\approx 2\) etiologies as defined above coexisted in the same individual; (7) unknown cause when no etiology was identified.\textsuperscript{44,45}

In view of the etiologic heterogeneity even of stroke subtypes such as lacunar infarcts,\textsuperscript{46} which account for about a quarter to one third of all ischemic strokes, it is questionable if such a subgroup should be “specified”. This procedure also raises the issue of correcting for multiple testing. Bevan and colleagues\textsuperscript{47} reported a significant association of a \(LTC4S\) promoter variant with small vessel disease but not with overall stroke. \(LTBR\) variants, which differed across study populations, were predominantly associated with cardioembolic stroke, which is mainly due to atrial fibrillation. Therefore, it could be argued that the genetic association identified with \(LTBR\) variants may be related to atrial fibrillation phenotypes rather than with stroke as such.

Along these lines, Kubo and colleagues\textsuperscript{48} reported on a significant association of the \(\text{protein kinase C (PKC) PRKCH}\) locus with lacunar infarction, but not atherothrombotic infarction—they wondered if this was not simply due to sample size—in a case-control study design. They subsequently replicated this association in independent samples from Biobank Japan and finally confirmed the association with lacunar infarction in a 14-year population-based follow-up study. The strength of the approach was to demonstrate robust “specific” association of the “second best-associated” variant \(\text{PKC}\eta \ V374I\) with lacunar infarction in different subsequent samples from retrospective-to-prospective design. However, authors did not report on functional differences at the mRNA and protein expression level for the different \(\text{PKC}\eta \ 374V\)- or \(374I\)-carrying constructs. Instead, authors proposed a potential difference in protein function in that \(\text{PKC}\eta \ 374I\) had 1.6 times higher activity than \(\text{PKC}\eta \ 374V\), which is not very convincing, as is also not the proposed minimal difference in autophosphorylation.

This should urge the scientific community to only claim causality (and herewith ‘effect’) for a genotype-phenotype relationship if functionality of a candidate variant is appropriately and convincingly confirmed in vitro and in vivo, usually in animal models.\textsuperscript{48}

For future purposes related to stroke genetics, different SNP (in k depending on the number of SNPs on the chip) sets are available. The first 100k set was launched in 2004,\textsuperscript{49} and included 116 204 SNPs, with a median and mean intermarker distance of 8.5 and 23.6 kb, respectively, with 92% of the genome within 100 kb of a SNP. In genome-wide association studies, 100k,\textsuperscript{14} 300k,\textsuperscript{15} and 500k sets\textsuperscript{50} have been proven useful in the independent identification and replication of associations of common variants on chromosome 9p21 with MI in white populations. All of these samples might have a
common underlying genetic imprint that might only be detected in large scale genomic applications. Besides the present versions of 500 or 1000 k chips, newly generated genotyping arrays in that range will cover additional gene regions in a complementary setting to already existing ones and continuously be upgraded by more present versions of human SNP maps (phase II HapMap, as of October 2007). As most of the SNPs included in the 100 k (and presumably in the 500 k set either) are not coding ones (60% represent SNPs 2 kb outside a gene; 30% of SNPs are intronic ones; <1% are coding variants), once a candidate gene locus is proposed, exhaustive resequencing is required to identify the responsible functional variant, even if this does not always warrant its/detection. Indeed, the ultimate goal of phenotype-driven approaches is to detect novel and unexpected biological pathways and networks that underlie the complex phenotype expression under study. It also offers new dimensions in necessitating the specific profiling of more complex molecular biological functionality of unexpected sequence regions, for example for remotely located as (for the ~400 bp long-range regulatory CNS-1 element on chromosome 5q31 that regulates interleukin-4, interleukin-13, and interleukin-5 gene expression, the latter one at a distance of ~120 kb) or intergenic regions (CNS-1 element is located intergenically as ~13 kb) between interleukin-4 and -13), that possess gene regulatory or gene product function capacities. Additional future research is needed to define the role of copy number variation (CNV) in CVD as new comprehensive CNV maps and detection technologies are available. In the advent of developing new technological and bioinformatics tools for deciphering complex genetic information at a quantitative population level, there is much hope to identify new targets for therapy and develop new strategies for prevention of CVD development and the occurrence of disease-related disabling or other fatal events.

Sources of Funding
S.-M. B.-H. is supported by a grant from the EU-Project Network of Excellence, FP6-2005-LIFESCIHEALTH-6, Integrating Genomics, Clinical Research and Care in Hypertension, InGenious HyperCare proposal No.: 037093.

Disclosures
None.

References


Where Do We Go for Atherothrombotic Disease Genetics?
Stefan-Martin Brand-Herrmann

*Stroke*. 2008;39:1070-1075; originally published online March 6, 2008;
doi: 10.1161/STROKEAHA.107.508515

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2008 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://stroke.ahajournals.org/content/39/4/1070

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published
in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office.
Once the online version of the published article for which permission is being requested is located, click
Request Permissions in the middle column of the Web page under Services. Further information about this
process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org//subscriptions/