Strategic Approaches to Unraveling Genetic Causes of Cardiovascular Diseases

A.J. Marian, John Belmont

Abstract: DNA sequence variants are major components of the “causal field” for virtually all medical phenotypes, whether single gene familial disorders or complex traits without a clear familial aggregation. The causal variants in single gene disorders are necessary and sufficient to impart large effects. In contrast, complex traits are attributable to a much more complicated network of contributory components that in aggregate increase the probability of disease. The conventional approach to identification of the causal variants for single gene disorders is genetic linkage. However, it does not offer sufficient resolution to map the causal genes in small families or sporadic cases. The approach to genetic studies of complex traits entails candidate gene or genome-wide association studies. Genome-wide association studies provide an unbiased survey of the effects of common genetic variants (common disease–common variant hypothesis). Genome-wide association studies have led to identification of a large number of alleles for various cardiovascular diseases. However, common alleles account for a relatively small fraction of the total heritability of the traits. Accordingly, the focus has shifted toward identification of rare variants that might impart larger effect sizes (rare variant–common disease hypothesis). This shift is made feasible by recent advances in massively parallel DNA sequencing platforms, which afford the opportunity to identify virtually all common as well as rare alleles in individuals. In this review, we discuss various strategies that are used to delineate the genetic contribution to medically important cardiovascular phenotypes, emphasizing the utility of the new deep sequencing approaches. (Circ Res. 2011;108:1252-1269.)

Key Words: complex traits • genetics • next-generation sequencing • polymorphism

The human nuclear genome (henceforth “genome”) is a simple and yet a complex structure. It is a large monotonous macromolecule comprising 3.2 billion repeating nucleotides of adenine (A), cytosine (C), guanine (G), and thymine (T), which are arranged in a seemingly random order. However, these four nucleotides not only determine expression of various biological and pathological phenotypes but also serve as the platform for various genomic and environ-
mental factors to exert their functional and biological effects. Consequently, elucidation of the molecular structure of the genome including its nucleotide sequence is fundamental to understanding the molecular pathogenesis of human diseases.

Sequencing of the human genome, however, has been a daunting task, at least until the very recent years. The Human Genome Project, which was launched in 1990 with the primary goal of deciphering sequence of the human genome, took more than a decade to complete, even in a draft form, and cost close to $3 billion.\(^1,^2\) DNA sequencing technology, however, has undergone a colossal shift during the past 6 years. Various new techniques that sequence millions of DNA strands in parallel have been developed. The new technologies, which are collectively referred to as the next-generation sequencing (NGS) platforms, as opposed to the Sanger method,\(^3\) which was used in the Human Genome Project, have increased DNA sequencing output and have reduced the cost of DNA sequencing by \(\approx 500,000\)-fold. These advances in DNA sequencing technologies along with the rapidly declining cost of sequencing are changing the approach to genetic studies of not only single gene disorders but also common complex disorders.

Despite its apparent simplicity, the genome is a complex structure. The complexity is far beyond the primary base sequence of the genome. DNA is a large macromolecule that requires a complex system to orchestrate its compaction inside the nucleus in a manner that selected genes are accessible to specific DNA processing enzymes, such as polymerases, in an orderly and dynamic fashion, as demanded by the cell in response to internal and external stimuli.\(^4\) Thus, understanding the functional content of the genome necessitates knowledge beyond the complete genome sequence. Based on today’s knowledge, only 1% of the human genome is transcribed into mRNA and translated into proteins. An additional 0.5% serves as a template for noncoding RNA and the regulatory regions that control gene expression.\(^5\) The functions of the remaining 98.5% of the genome including functional conserved noncoding elements, which comprise at least 6% of the genome,\(^6\) remain unknown. Hence, this large segment of the genome is referred to as the dark matter of the genome.\(^5\) The discoveries of noncoding RNA, microRNA, splice variants, and regulatory elements in trans point to the complex mechanisms by which the genome governs various biological processes, including phenotypic expression of diseases (Figure 1). To elucidate the determinants of any biological and clinical phenotype, a comprehensive approach that not only utilizes information content of the nucleotides sequence but also utilizes that of the transcripts, whether coding or noncoding, chromatin structure and function, and transcriptional machinery that orchestrates gene expression, among the others, would be necessary. The focus of this review is on strategic approaches to identify the DNA sequence variants (DSVs) that either strongly determine disease risk, as in single gene disorders, or influence susceptibility to a disease. Because the majority of the known disease-causing DSVs are located within exons, the current focus of human genetic studies is on whole exome sequencing.\(^7\) With advances in DNA sequencing technology and increasing knowledge of the nonprotein coding regions of the genome, one expects a rapid shift from whole exome to whole genome sequencing as the desirable approach to identify disease-causing or disease-associated DSVs.

**Genetic Diversity**

Catalogs of common genetic variation have been accumulated over the past three decades. For example, the database of single nucleotide polymorphisms (dbSNP Build 132) has \(\approx 37\) million entries. The number of polymorphic variants in a single genome, however, was largely unknown until the report by Venter\(^8\) of the diploid genome sequence in 2007. The findings were notable for presence of 4.1 million DSV, including \(\approx 3.5\) million single-nucleotide polymorphisms (SNPs), which affected \(\approx 44\%\) of the annotated genes. Likewise, the Venter genome contained \(\approx 10,000\) nonsynonymous SNPs (nsSNPs), of which \(\approx 7000\) were considered potentially harmful variants. In addition, structural variations (SV), which involved up to several million nucleotides, comprised three-quarters of the variant nucleotides.

Subsequent sequencing of additional individual genomes confirmed the findings of extensive DSV in each genome and pointed out the presence of a large number of novel variants. Today, the genome and exome sequences of a relatively large number of individuals have become available, along with the HapMap and 1000 genomes data.\(^9,^15\) The initial results of the 1000 genome projects indicate that each genome has \(\approx 250\) to 300 loss-of-function variants in the annotated genes and 50 to 100 variants that already have been implicated in inherited disorders.\(^10^\) In addition, each genome has \(\approx 30\) de novo variants, a finding that indicates a germline mutation rate of \(1 \times 10^{-8}\) per generation.\(^16\) Furthermore, each genome has several large (>50 Kbp) and \(\approx 100\) heterozygous copy number variants (CNVs) covering \(\approx 3\) Mbp.\(^17,^18\) Collectively, the data indicate that humans differ in \(\approx 0.12\%\) of their genomes, or \(\approx 4\) million DSV per genome, comprising \(\approx 3.5\) million SNP and several hundred thousand structural variations including CNV.

---

**Non-standard Abbreviations and Acronyms**

- **CNV**: copy number variants
- **dbSNP**: SNP database
- **DSV**: DNA sequence variants
- **Gb**: gigabases
- **GWAS**: genome-wide association studies
- **Indel**: insertion/deletion
- **Kbp**: kilo basepair
- **LD**: linkage disequilibrium
- **LDL-C**: low-density lipoprotein cholesterol
- **MAF**: minor allele frequencies
- **Mb**: million basepair
- **NGS**: next-generation sequencing
- **nsSNP**: Nonsynonymous single nucleotide polymorphisms
- **SiRNA**: short-interfering RNA
- **SNP**: single nucleotide polymorphism
The functional and biological significance of the vast number of DSV in the human genome are unknown. Nevertheless, they are expected to exert effects that follow a gradient ranging from negligible to severe. Among the 10,000 nsSNPs in each genome, approximately two-thirds are predicted by in silico analysis to be deleterious to function. Likewise, structural variations that encompass several thousand to million base pairs could duplicate or delete a gene or multiple genes and, hence, would be expected to hold significant clinical implications. Nevertheless, in a given clinical phenotype, a small number of alleles are expected to exert large effects, a few moderate effects, and a very large number of modest or no effects. Presumably, clinical phenotype is the consequence of the additive effects and interactions among multiple alleles with varying magnitude of effect.

**Genetic Mechanisms of Human Diseases**

**Common Disease–Common Variant Hypothesis**

Common cardiovascular diseases have considerable genetic components, as evidenced by familial aggregation and twin studies. The estimated heritability of common complex diseases, defined as a proportion of the phenotypic variance accounted for by genetic factors, varies from 20% to 80%, depending on the phenotype and study characteristics. In contrast to single gene disorders, wherein a single DSV imparts a large determinative effect, no single allele or locus dominates as the determinant of a complex phenotype. Accordingly, complex diseases result from the cumulative and interactive effects of a large number of loci, each imparting a modest marginal effect on expression of the phenotype (Figure 2). The common disease–common variant hypothesis posits that multiple common alleles, defined as alleles with a population frequency of $0.05$, contribute to the risk of common diseases developing. The common disease–common variant hypothesis underpins genome-wide association studies (GWAS), wherein cases and controls are genotyped for hundreds of thousands of common variants. In GWAS, linkage disequilibrium (LD), the correlation between markers, is exploited to tag common variants that influence medically important traits. Effective tag SNP and their underlying haplotypes in selected reference populations have been completed for two model organisms. (Illustration Credit: Cosmocyte/Ben Smith).
the attributable risk of a common allele in a population might be considerable, simply because of its high minor allele frequency (MAF).

Rare Variant–Common Disease Hypothesis

Although GWAS have been successful for identifying many loci associated with important cardiovascular diseases, there are additional large gaps in our understanding of the genetic contribution to these conditions. Common alleles, at least from the perspective of their individual marginal effects, account for a relatively small fraction of the total heritability of those disorders. As such, SNPs identified in GWAS of systemic hypertension, dyslipidemia, and cardiac conduction intervals account for only a small fraction of interindividual variance. However, the attributable risk of a common allele in a population might be considerable, simply because of its high minor allele frequency (MAF).

Gradients of Allele Frequencies and Effects

The shortcomings of GWAS in explaining the heritability of common complex disease often referred to as the “missing heritability” might be in part because of the presence of rare DSV with relatively large effects that are not tagged by the typical marker sets used in GWAS. Rare alleles are typically defined as those that have a MAF ≤0.01 in a population. Whenever rare DSV are unique to an individual or to a family, they are considered as “private” variants. Possible contributions of rare alleles with large effects on common diseases have led to the rare variant–common disease hypothesis. In support of this hypothesis, uncommon and rare alleles in genes known to cause single gene disorders could contribute to susceptibility for complex phenotypes and enhance detection of otherwise clinically undiagnosed conditions. A notable example is identification of mutations in MYH7, MYBPC3, TNNI3, and MYL3 genes, causal genes for autosomal–dominant hypertrophic cardiomyopathy, in a subgroup of individuals in a community-based cohort who had an increased left ventricular wall thickness. Similarly, multiple rare alleles in ABCA1, the responsible gene for Tangier disease, also contribute to plasma high-density lipoprotein cholesterol levels in the general population. Likewise, multiple rare alleles in genes not associated with single gene disorders might account for a substantial portion of heritability of complex traits. The notion that a number of rare variants might impart large effects on the phenotype is plausible. However, currently, there are insufficient data to substantiate this hypothesis and conclude that multiple rare alleles are major components of missing heritability.

Figure 2. Gradients of disease prevalence, minor allele frequencies, and effect sizes. The prevalence of disease, number of genetic determinants, and the effect sizes of the DNA sequence variants are depicted as continuums. Single gene disorders are caused by rare variants with large effect sizes. Typically, several other variants also expected to contribute to phenotypic expression of the diseases. On the opposite end of the spectrum are the common complex traits, which are caused by a very large number of genetic variants, each imparting a modest effect size.
The more tightly coupled the genetic variant, in terms of biological functions, such as enzymatic activity or protein interactions, the greater the effect. Likewise, a small number of variants with relatively large effects are observed to influence mRNA expression levels of their respective genes. Therefore, genetic analysis of biomarkers, including transcript abundance, can be a very effective strategy for “divide and conquer” of the much more complex origins of cardiovascular disease.

Approach to Genetic Studies

Genetic factors are components of the “causal field” for virtually all medically important traits. Causal fields can be composed of necessary and sufficient factors, as in the case of single gene disorders. Most often, however, there is a much more complicated network of unnecessary and insufficient components that in aggregate increase the probability of disease. These factors are contributory causes; they are nonredundant components of pathways that by themselves may be unnecessary but if altered can influence the occurrence of disease. The genotypic effect can be represented as a probability from 0 to 1 that it influences a particular clinically important trait. Genotypes with large effects are best represented by single gene disorders, such as familial cardiomyopathies wherein a single mutation leads inexorably to a major disease (Figure 2). Even in the single gene disorders, the severity of the disease is influenced not only by the causal mutation but also by genetic modifiers and environmental factors. On the other end of spectrum of genetic effects are much weaker effects, as uncovered in GWAS of complex phenotypes such as atherosclerosis (Figure 2).

Candidate Gene Association Studies

The approach is based on previous knowledge of candidacy of the gene(s) of interest in the pathogenesis of the phenotype. Unless performed in a very large sample size of well-characterized populations, the approach is prone to spurious results, particularly for distant phenotypes. The approach has led to identification of a number of alleles that influence responsive to drugs, both in terms of efficacy as toxicity (pharmacogenetics). For example, DSV in CYP2C9 and VKORC1, genes encoding cytochrome P450 isofrom 2C9 and vitamin K epoxide reductase, respectively, are associated with response to treatment with anticoagulant coumadin. Similarly, DSV in genes coding for P-450 enzymes CYP3A4, CYP3A5, and CYP2C19 are associated with responsive to treatment with antiplatelet agent clopidogrel, a prodrug that is converted in the liver to an active metabolite. Likewise, DSVs in α2C-adrenergic and β1-adrenergic receptors are associated with the response of patients with systolic heart failure to treatment with β-blockers. Moreover, DSVs in APOE, PCSK9, and HMGCR have been implicated in response to statins. Regarding drug toxicity, DSV in SLC01B1 encoding solute carrier organic anion transporter 1B1 is associated with statin-induced myopathy. Likewise, DSVs in genes causing congenital long-QT syndrome are associated with drug-induced cardiac arrhythmias. DSVs implicated in pharmacogenetics appear to have moderate effect sizes and, hence, might have some clinical implications in guiding drug efficacy and avoiding toxicity (Figure 3).

GWAS

GWAS have been widely used to delineate the genetic basis of common complex disorders. GWAS are case-control studies, wherein research subjects are typed for a large number of SNP, typically 300 000 to 1 000 000 SNP/CNV, and the allele or genotype frequencies are evaluated for differences between groups or for correlations with continuous traits. GWAS is primarily designed to provide an unbiased survey of the effects of common genetic variants. Markers chosen for GWAS typically have MAF of ≥0.05 and are selected to “tag” the most common haplotypes observed in the major continental populations. Such tagging is more complete in European and East Asian populations compared to African populations because of inherent differences in the LD patterns.

The power of the GWAS to detect the phenotype-associated alleles depends directly on the sample size of the study population, MAF, strength of LD between the markers, and the causal variants and the effect sizes of the alleles. The density of the genotyping arrays have increased significantly over time and the current versions can easily genotype as many as 2.5 million SNPs and CNVs. Collectively, these advances have enhanced the power to detect the associated alleles. During the past decade, GWAS have been completed for a very large number of cardiovascular phenotypes. The National Human Genome Research Institute maintains a catalog of published GWAS that can be accessed at http://www.genome.gov/gwastudies/.
A major strength of GWAS is that they may lead to identification of novel pathways involved in the pathogenesis of the phenotype. Despite the apparent simplicity, however, the results of GWAS are subject to multiple-hypothesis testing because of typing of a very large number of SNPs and, hence, beget correction for the possibility of a random association attributable to multiple testing. In addition, often the same study population is analyzed for the association of the genotypes with multiple phenotypes, which also increases the likelihood of spurious associations. Therefore, statistical corrections for multiple hypotheses testing are essential. The best approach to correct for multiple hypotheses testing in GWAS remains to be established. The conventional Bonferroni method for correcting for multiple testing \((P = \alpha/n)\) is considered too conservative because the Bonferroni correction assumes that the tests are independent, which is not the case for GWAS markers because of residual LD in local regions of the genome. Permutation tests are probably the most robust for correcting for multiple testing but are computationally very intensive and impractical considering the very large number of genotypes \((\text{SNP} \times \text{individuals})\) in the GWAS. Various statistical methods have been applied to correct for multiple testing in GWAS and determine the threshold for statistical significance. Based on these calculations, \(P < 5 \times 10^{-8}\) or more stringently \(P < 1 \times 10^{-8}\) is considered evidence of a strong association.

GWAS have been extremely useful for identifying a very large number of phenotype-associated alleles, including many novel loci. The paucity of discovering functional SNP as the associated alleles, however, is notable. Accordingly, the results of GWAS have typically not been fruitful in immediate elucidation of the responsible mechanisms behind the observed genetic association. Consequently, GWAS demand complementation with robust mechanistic studies to elucidate the biological mechanisms responsible for the genetic association.

Alleles identified in GWAS are seldom the true causative alleles but are likely in LD with the true causative alleles. Thus, extensive additional studies are typically required to complement the results of GWAS to identify the disease-causing alleles. These shortcomings render the results of GWAS in the discovery population as provisional, requiring replications in independent study populations and ultimately validation through experimentation.

The results of GWAS have minimal to modest impact, if any, on preclinical diagnosis, risk stratification, or genetic-based prevention and treatment at an individual level. Therefore, the significance of additional mechanistic studies cannot be overemphasized. However, the appropriate platforms to validate the results of GWAS through molecular mechanistic studies remain to be established. The challenge is best illustrated for delineation of the responsible mechanisms for the observed association of SNP at 1p13 locus with plasma cholesterol levels and coronary atherosclerosis.46–48 Accordingly, the minor allele of the rs599839 SNP located in \(SORT1\) gene at the 1p13 locus is associated with a decrease in serum low-density lipoprotein-cholesterol \((\text{LDL}-\text{C})\) by 0.14 mmol/L and a 9% decrease in risk of coronary atherosclerosis.47 Two recent studies attempted to elucidate the responsible mechanisms but, unfortunately, reported discordant results.47,49 Musunuru et al fine-mapped the locus and on further analysis defined rs12740374 to be the causative SNP. The minor allele of this SNP created a C/EBP-\(\alpha\) binding site on \(SORT1\) promoter, which enhanced transcriptional activity. Overexpression of \(SORT1\) in mouse liver using recombinant adeno-associated viruses led to reduced plasma LDL-C levels. In contrast, siRNA-mediated suppression of expression of \(SORT1\) in the liver had the opposite effect. Therefore, the studies identified increased expression level of \(SORT1\) imparted by the minor allele of rs12740374 as the responsible mechanism for reduced plasma LDL-C level.49 However, studies by Kjolby et al showed the opposite effects. Accordingly, overexpression of \(SORT1\) stimulated hepatic release of lipoproteins and led to increased plasma LDL levels.49 Thus, despite the concordant and reproducible results of GWAS, studies to delineate the responsible mechanisms have not led to concordant results. The contrasting results may hint to the challenges encountered in recapitulating the results of human genetic studies of complex phenotypes in model organisms.

Identification of the responsible mechanisms for association of the SNP at the 9p21 locus, which has been robustly linked to atherosclerosis,51–53 also has been challenging. The refined locus does not contain a known gene, but the region contains cyclin-dependent kinase inhibitor 2A and 2B \((\text{CDKN2A}, \text{CDKN2B})\), methylthioadenosine phosphorylase \((\text{MTAP})\), and \(ANRIL\), the latter codes a long noncoding RNA. None of these genes appears to be a biologically plausible candidate gene for atherosclerosis. The cell-cycle regulators \(\text{CDKN2A}\) and \(\text{CDKN2B}\) are tumor suppressor proteins and markers of cell senescence.54 Deletion of the orthologous region of 9p21 locus in the mouse genome is associated with reduced expression levels of \(\text{Cdkn2a}\) and \(\text{Cdkn2b}\), enhanced proliferation, and reduced senescence of smooth muscle cells.55 The findings suggest accelerated smooth muscle cells proliferation as the potential mechanism for the observed association of the 9p21 locus and atherosclerosis. \(\text{MTAP}\), which is also located at the region, codes for an enzyme that is involved in polyamine metabolism and generation of adenine and methionine. Deletion of this gene is embryonically lethal and in homozygous form is associated with reduced life span because of severe lymphoproliferative disease resembling T-cell lymphoma.56 The third gene is \(\text{ANRIL}\), which is expressed in cells involved in atherosclerosis, such as smooth muscle cells, endothelial cells, and macrophages. It has multiple isoforms, but none has an open reading frame and, hence, the gene does not appear to code for a protein. Deletion of the 9p21 region involving \(\text{ANRIL}\) is implicated in melanoma and solid tumors.57

The 9p21 locus, despite being a gene desert, is extremely rich in enhancers.58 It contains at least 33 enhancers, including one that interacts with two risk alleles for coronary artery disease at this locus. The risk alleles disrupt the binding site for STAT1, which is the signal transducer for a variety of ligands including interferon-\(\alpha\), interferon-\(\gamma\), and cytokines. Binding of STAT1 to the wild-type alleles inhibits expression of \(\text{CDKN2B-AS}\) (nonprotein coding \(\text{CDKN2B}\) antisense RNA1).58 Treatment of endothelial cells carrying the risk
Table 1. Potential Explanations for the Modest Capture of Heritability by Genome-Wide Association Studies

<table>
<thead>
<tr>
<th>Explanation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A large number of common variants with low magnitude of effect</td>
<td>Polygenic inheritance</td>
</tr>
<tr>
<td>Rare variants with large effects</td>
<td>Single gene or oligogenic inheritance</td>
</tr>
<tr>
<td>Structural variants</td>
<td>Rare variants with high mutation rates</td>
</tr>
<tr>
<td>Interactions between alleles</td>
<td>At homologous loci (dominance) and between alleles at non-homologous loci</td>
</tr>
<tr>
<td>Parent of origin effects</td>
<td>Epigenetics effects</td>
</tr>
<tr>
<td>Underestimation of the effect of shared environment among relatives leading to inflated estimate of heritability</td>
<td></td>
</tr>
</tbody>
</table>

The problem of “missing heritability” has raised significant concern about the utility of GWAS in delineating the genetic basis of complex traits. The “missing heritability” in part may reflect the definition of heritability that in narrow sense is defined as the proportion of the phenotypic variance attributable to additive genetic factors. The definition reflects a general model of gene action that has only coarse explanatory power and is subject to many difficulties of estimation. In fact, Feldman and Lewontin anticipated most of the problems related to heritability interpretation in complex disease long before the current era of GWAS. Some of the possibilities that may account for the relatively modest portion of the heritability captured by GWAS are listed in Table 1.

In the rare variant–common disease hypothesis population, genetics theory predicts and empirical observation demonstrates that there are large numbers of rare alleles, among which some have large effects. There are many fewer common alleles with weak effects and it is these alleles that are reliably identified in GWAS. However, if one counts the absolute number of people with each kind of allele, the relationship is inverted. For any particular locus there are small numbers of people with rare alleles and large numbers with common alleles. For this reason, the attributable fraction for a particular locus can be dominated by the common alleles, even though the rare alleles have much larger effects. One way to consider the attributable fraction is the fraction of cases that would be eliminated if the allele were not present. Therefore, even weakly acting common alleles...
can be the most significant contributors to cardiovascular disease.

**Direct DNA Sequencing**

The cost of sequencing the entire human genome is expected to decrease to $\approx 1000$ by the end of 2011. This evolution has been made possible by switching to massively parallel sequencing platforms wherein millions of DNA strands are sequenced in parallel and simultaneously. The technologies have made it feasible to sequence two or three genomes or a dozen of exons in a week. A major advantage of the whole genome and exome sequencing approaches is in its enabling principle that allows detection of not only the common (MAF $>0.05$) and uncommon (MAF $<0.05$ $\rightarrow 0.01$) but also rare (MAF $<0.01$) and private (found only in the probands or genetically related immediate family members) variants.

Application of the NGS extends beyond the DNA sequencing because the core genome technology also affords the opportunity to sequence and analyze the whole transcriptome (RNA-Seq), epigenetic modifications (Methyl-Seq), and transcription factor binding sites (ChIP-Seq). The approach is quantitative and enables relatively small amount of template. In the present review, the focus is on DSV.

**Next-Generation Sequencing Platforms**

Sydney Brenner, Nobel Laureate in Physiology and Medicine (2002), introduced the first technique of sequencing of millions of copies of the DNA simultaneously, referred to as MPSS in 2000. Soon, George Church et al described the technique of multiplex polony sequencing. The first commercial NGS platform was based on pyrosequencing technique. However, it was soon surpassed in output by reversible dye termination and sequencing by ligation approaches. Sequencing platforms continue to evolve at a rapid pace with enhanced capacity to generate bigger outputs and more accurate reads. Accordingly, the newer instruments can generate up to 300 Gb of throughput per sequencing run, which would be sufficient to cover two to three genomes and approximately a dozen exomes and transcriptomes. Detailed technical review of the existing platforms is beyond the scope of the present review and can be found elsewhere. The two most commonly used platforms for whole exome and whole genome sequencing are the SOLiD systems (Applied Biosystems), which are based on sequencing by ligation-based chemistry and HiSeq systems (Illumina), which utilize reversible terminator-based sequencing by synthesis chemistry. Both platforms generate short reads that typically are 50 to 120 bases long and each can generate $\approx 20$ to 30 Gb per day. The accuracy of the sequence reads depends on various factors, including depth of coverage. Overall, the systems have a high accuracy rate, typically $>99.9\%$. However, given the vast size of the sequence output, even a very low error rate can lead to a considerable number of erroneous calls and downstream work. For medical sequencing, it is essential to validate the variant calls either by an alternative method, such as Sanger sequencing, or by repeating the deep sequencing in toto and accepting only those variants that were reproduced.

In contrast to short-read NGS platforms, pyrosequencing (Roche 454 sequencing systems) can generate a read length of $\approx 400$ bases and $>1$ million reads per run in $\approx 10$ hours. However, the size of sequence output is much smaller and the cost per base is much higher. Because of the length of the reads, the system is best suited for de novo sequencing. The error rate is $\approx 0.1\%$. Therefore, for medical sequencing, confirmation of the variants is essential.

Newer techniques include single molecule real-time sequencing, which is also referred to third-generation sequencing, can generate an average read length of $>1000$ bases. However, the system at the present time has a high error rate and does not seem to be suitable for medical sequencing. Finally, NGS platforms are also available for sequencing of small genomes and targeted sequencing of relatively small regions or small number of genes.

**Whole Genome Sequencing**

Whole genome sequencing using NGS instruments only recently has become feasible in individual laboratories. The existing platforms afford the opportunity to sequence one to three genomes in a single run in 7 to 8 days. However, currently, only few centers have the sequencing and bioinformatics capacity and financial means to handle large-scale whole genome sequencing projects. Technical aspects include the size of the mappable data (to the reference sequence), depth of coverage, error rate of allele calling, and the gaps in the coverage. Technical advances have made it feasible to apply whole genome sequencing to identify the genetic cause of Mendelian disorders, at least in proof-of-principle studies. Application of this approach to other single gene and multigene disorders is likely to accelerate significantly during the next few years. The advantage of whole genome sequencing is that it affords the opportunity to detect all SNP, whether coding or noncoding, and to some extent CNV in the genome. It also does not depend on a target capture technology, which may suffer from unequal capture of the desired genomic regions. The disadvantages are simply the limited capacity of most laboratories to handle and store terabytes of data that are generated by the sequencer and bioinformatics. Various algorithms have been developed to restrict the number of candidate variants and facilitate identification of the causative variants. The key components are population MAF (novel, known, de novo), type of the variants (deletion, frame shift, missense, splice), evolutionary conservation of the variants, and expected biological effects.

**Whole Exome Sequencing**

The whole exome sequencing approach is designed to capture, enrich, and sequence all exons in the genome. Each genome is estimated to contain $\approx 300$ Mbp representing $\approx 180000$ exons of $\approx 23000$ protein-coding genes. The focus on whole exome sequencing as opposed to whole genome sequencing stems from the existing data, which indicate that more than two-thirds of the known disease-causing genes in humans are located within exons. Steps involved in whole exome or subgenomic sequencing include library preparation, target capture, target enrichment, and sequencing. Commercially available capture technologies enable efficient capture of the exome and their sequencing. Nonetheless, the efficiency of capture could vary in different genomic regions and
often 5% to 20% of the exons may not be captured and sequenced adequately to afford robust allele calling. Whole exome sequencing has been successfully applied to identify the genetic causes of rare Mendelian disorders such as Freeman-Sheldon syndrome, congenital chloride-losing enteropathy, Kabuki syndromes, and hypertension caused by hyperaldosteronism. Whole exome sequencing to identify genetic causes of uncommon and heterogeneous Mendelian disorders could face formidable challenges, particularly in small families and sporadic cases, to discern the disease-causing variants from those that by chance alone are presented in the affected individuals.

Targeted Subgenomic Sequencing

Targeted subgenomic sequencing is in essence similar to whole exome sequencing except that selected exons or subgenomic regions are amplified by long-range polymerase chain reaction or captured using custom-made capture probes, enriched and sequenced using NGS platforms. The approach might be desirable for genetic screening through long-range polymerase chain reaction or capture and subsequent sequence of all exons in the known genes implicated in the phenotype, such as screening of the known genes coding for sarcomeric proteins in patients with cardiomyopathies and their family members. This approach is not much less expensive than the whole exome sequencing approach but is clearly less demanding in terms of bioinformatics. Like all target capture and sequencing approaches, it has the problem of uneven capture and polymerase chain reaction amplification of the intended targets and the risk of underdetection. Likewise, the approach by definition is limited to known targets and, therefore, somewhat is limited in its scope. Moreover, given the feasibility and declining cost of whole exome/genome sequencing, and in view of the complexity of the genetic determinants of the clinical phenotypes, subgenomic sequencing is best suitable for specific circumstances, such as follow-up studies to genetic linkage and GWAS.

Design of Genetic Studies

A phenotype is in part the consequence of effects of multiple common and rare alleles each imparting a gradient of effects. Whereas GWAS is typically designed to identify common alleles, the NGS approach is best suited for a comprehensive detection of common as well as rare variants. The power of NGS to identify the causal variants is primarily determined by design of the study and the characteristics of the population. As in all genetic studies, robust phenotyping and family-based genetic studies are far superior to studies in isolated cases or in a cohort of sporadic cases.

Phenotyping

Robust phenotyping is an essential but often an inadequately defined component of the genetic studies. Clinical phenotyping often does not offer sufficient resolution or specificity. Even the most clinically robust phenotype, such as all-cause mortality, is subject to enormous etiologic heterogeneity. However, discerning the etiologic subtypes renders the ap-
proach to uncertainties of accurate identification of the subtypes. Likewise, phenotypic admixture is also not uncommon, as illustrated for the commonly pooled phenotypes of coronary atherosclerosis, ischemic heart disease, and myocardial infarction, as a single phenotype. Whereas these phenotypes have overlapping components, each has partially separate mechanistic basis. Likewise, clinical phenotypes are usually a continuum but often are considered categorical. For example, dichotomization of coronary atherosclerosis, a continuous phenotype, as a categorical phenotype of <60% or >60% minimum lumen diameter stenosis is not only subject to the imprecision of the quantification, which could be quite large, but also inadequately represents the phenotypic burden. Phenocopy conditions also compound and confound accurate diagnosis. Collectively, the inadequacies of accurate clinical phenotyping reduce the successful elucidation of the genetic basis of various clinical phenotypes.

Phenotypic plasticity of mutations in a given gene also complicates a straightforward genotype–phenotype correlation. This is most remarkable for single gene disorders and best illustrated for LMNA, which encodes Lamin A/C, an important component of the inner nuclear lamina.83 A diverse array of mutations in LMNA cause at least 13 distinct phenotypes, which are collectively referred to as laminopathies.84 Phenotypic expression of LMNA mutations in the heart or cardiolaminopathies is notable for dilated cardiomyopathy, supraventricular bradyarrhythmias, and conduction defects.85 Likewise, mutations in genes coding for sarcomeric proteins, such as MYH7 and TNNT2, can cause either dilated cardiomyopathy or hypertrophic cardiomyopathy, which are on the opposite ends of phenotypic spectrum of cardiac responses to mutations or nongenetic factors.86–88 Similarly, mutations in SCN5A, which code for a sodium channel, are phenotypically expressed as the long-QT syndrome, Brugada syndrome, atrioventricular conduction defects, atrial fibrillation, and dilated cardiomyopathy.89–91 Phenotypic plasticity appears to be the expected rather than the exception for various DSVs in the same gene, but not typically for the same DSV in a given gene.

Biological variability and shortcomings of the quantification methods also diminish the power to map genetic determinants of certain clinical phenotypes. The simplest example is the measurement of systolic and diastolic blood pressure values utilizing a sphygmomanometer, which is based on the detection of Korotkoff sounds. A single measurement of blood pressure is often inadequate and seldom two measurements even when measured within a short time period are identical. Likewise, biochemical phenotypes, such as plasma levels of proinflammatory cytokines and C-reactive protein, are quite dynamic and exhibit considerable intraindividual variability. A single measurement is usually inadequate to reflect the physiological or pathological burden of the phenotype. Physiological and technical variability are typically handled by increasing the sample size of the study population, which increases the power to detect significant effects. Nonetheless, increasing the sample size not only increases the cost but also renders the relevance of the findings to a single individual remote. Imperfectness of phenotyping is probably partially responsible for the “missing heritability” in the genetic studies of complex traits, because the identified DSV only account for a small fraction of the heritability.

Family-Based Studies
Family studies provide the most robust approach for delineation of the genetic determinants of the phenotype. Diseases with the strongest familial inheritance are single gene disorders, which are uncommon and often rare. Therefore, the number of DSVs in each genome with very large effects is also expected to be low. According to 1000 genomes data, each genome encompasses 250 to 300 loss-of-function variants in the annotated genes, 50 to 100 variants that already have been implicated in inherited disorders, and 30 de novo variants.92 DSV with large effects are easier to identify and establish as determinants of a phenotype, as illustrated in familial single gene disorders. Deep sequencing approaches are expected to supplant microarray-based genotyping approaches for identification of genetic determinants of single gene disorders. Likewise, the approach might enable identification of the DSVs with moderate effects that may serve as modifier alleles in single gene disorders. The significance of the latter is noteworthy because of the influence of genetic background, namely modifier alleles, on phenotypic expression of single gene disorders.92

Rare DSVs are also expected to contribute to phenotypic expression of common complex phenotypes, which show a familial aggregation. The stronger the evidence for a familial aggregation of a complex phenotype, the more likely is the presence of rare variants with large effects. In general, a larger number of genetically related family members provide a greater power to identify the causal and modifier variants, regardless of the approach being genotyping or deep sequencing. Likewise, family-based deep sequencing studies are most powerful when the causal variants occur de novo or are very rare. Moreover, deep sequencing is more powerful for identification of genetic causes of rare than common disorders. It is also more power for identification of genetic cases of Mendelian disorders with a recessive than those with an autosomal-dominant mode of inheritance. Nevertheless, deep DNA sequencing could enable elucidation of genetic basis of single gene disorders in small families and probands, wherein the conventional genetic linkage studies are not sufficiently powerful to map the chromosomal locus. The powerful of whole genome/exome sequencing to pinpoint the causal DSV also inversely correlates with prevalence and genetic heterogeneity of the disorder. In relatively common single gene disorders, a large number of alleles are expected to cosegregate with the phenotype and discerning the true causal variants from those that segregate with the phenotype by chance alone is challenging. Moreover, unlike the candidate gene approach, which is based on a priori knowledge, whole genome/exome sequencing is free of a priori assumption. The approach has been successfully applied to identify the causal variants in rare autosomal recessive or autosomal dominant diseases, such as Kabuki syndrome, Miller syndrome, and hyperaldosteronism.77,78,80,93

Although it is desirable to sequence the genome/exome in all related family members, the approach is currently costly and the analysis is demanding. An alternative approach is to
focus on family members who are more distantly related but are phenotypically affected, and as such family members are expected to share a lower number of alleles. The approach by reducing the number of shared alleles is expected to enhance identification of the causal variants. Another desirable and practical approach is to sequence and contrast sequence data from family members who are on the opposite ends of the phenotypic spectrum, for example, mild vs severe phenotype. Similarly, the approach is expected to enrich the chance of identifying genetic variants that impart relatively large effect sizes. Nonetheless, the challenge of identifying the causal variants is magnified inversely with the size of the families.

**Trio-Based Family Studies**
A trio in a family study refers to parents and an offspring. A deep sequencing strategy in a single trio does not offer much power except for the detection of rare and de novo variants in a biologically plausible or previously implicated gene in an affected offspring. However, sequencing of a large number of trios could afford the opportunity to apply the transmission disequilibrium test to identify the putative causal variants. Transmission disequilibrium test assesses inheritance of an allele by an affected offspring from an affected parent. In a case of no association, it is random event and there is a 50% chance of association. In the case of an association, the frequency of transmission deviates significantly from the chance.

**Sporadic Cases**
Application of deep sequencing technologies in sporadic cases requires a case-control study design similar to those conducted in GWAS. Unlike GWAS, however, deep sequencing will identify rare as well as common alleles and the frequencies of the alleles are compared in cases and controls in an allele-centric or gene-centric (collapsing) approach to test for the presence of statistically significant differences. The design of a case-control study in a deep sequencing project is of utmost importance because an ill-conceived study design could lead to identification of an exceedingly large number of DSV that differ between the cases and controls. Extensive amount of downstream analyses and experiments would then be required to discern the true associations from false. Several strategies could be utilized to strengthen the design of the case-control studies and reduce the number of putative candidate variants for subsequent validation. One such approach is to focus on cases that exhibit the extreme ends of the phenotype of interest (for example, those with severe and premature disease). Likewise, it is often desirable to include a group of “super normal controls,” which have been exquisitely phenotyped to exclude potential subclinical phenotype and have no family history of the phenotype of interest. Moreover, inclusion of the cases that have been enriched for the genetic load, such as an inbred population, as well as prioritizing of analysis of variants that are located in the previously mapped GWAS loci for the phenotype of interest could increase the likelihood of restricting the number of putative candidates.

**Bioinformatics and Statistical Analysis**
Various filtering algorithms are applied to restrict the number of putative candidates and to identify the causal variants among the myriads of alleles identified through deep sequencing. The field of bioinformatics is rapidly evolving and considerable progress has been made in eliminating the current bottleneck in analysis of the NGS data. The approach to identify the causative alleles is logical and based on the family structure (cosegregation), zygosity, novelty, being de novo, MAF, evolutionary conservation, and known or anticipated biological effects of the variants. Table 2 lists the most likely putative disease-causing variants.

The process typically involves mapping the sequence readouts to the reference sequence, which is successfully achieved for more than half of the reads. Using the mappable sequence, various bioinformatics programs are used to identify single nucleotide variants, small indels, and even CNV and inversions, depending on the sequencing platform. The accuracy of allele calling depends in part on the coverage depth and quality of the reads. As would be expected, a higher coverage depth would be required for calling heterozygous than homozygous variants. Although the bioinformatics programs are evolving rapidly, various software are already available to annotate the alleles in terms of quality of the call, coverage statistics, novelty or known frequencies of the variants, type of the variants, and their putative functional effects. A partial list of bioinformatics programs and their main applications are shown in Table 3.

Confirmation that a locus contributes to a disease must be based on statistical support and, ultimately, replication of the finding in independent cohorts of cases. There are now a few statistical methods that have been specifically tailored to address the comparison of rare variants between cases and controls. In 2008, Leal et al described the combined multivariate and collapsing method. Combined multivariate and collapsing combines collapsing of rare variants into a single class and multiple-marker tests for common variants.

**Table 2. High-Priority Variants Identified After Streamlining of Deep Sequencing Output**

<table>
<thead>
<tr>
<th>Type of the variants</th>
<th>Known disease-causing variants</th>
<th>Novel variants in genes known to cause the phenotype</th>
<th>Novel variants in the class of genes known to cause the phenotype</th>
<th>Novel variants in genes not previously not implicated in the pathogenesis of the phenotype</th>
<th>De novo variants that co-segregate with the phenotype in subsequent generations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
and has much greater power than single marker tests. In a recent innovation, Leal et al have described the kernel-based adaptive cluster method, which directly addresses the problem of detecting rare variant associations in the presence of functional misclassification. The sample risk is modeled using a mixture distribution with two components, noncausal and causal. The method uses continuous adaptive weighting in the comparison between cases and controls. As reference data sets become larger over the next few years, it should be possible to apply even more sophisticated methods that model the differences in mutation rate between loci and the known biological and functional significance of DSVs, but also inpast several decades are primarily based on phenotype-based approaches. The imperfectness of clinical phenotyping begets a shift toward using surrogate phenotypes that are proximal to genes and more likely to be subjected to larger effects (Figure 3). The significance of finding genetic determinants of the proximal phenotypes is that it not only could elucidate biological and functional significance of DSVs, but also might translate and extend to clinical phenotype. For example, heterozygous loss-of-function mutations in PCSK9 lower plasma LDL-C levels and reduce the risk of coronary heart disease drastically over a 15-year period. However, a discord between association of a biochemical (proximal) and clinical phenotypes with DSVs might be present. The discord is particularly evident when the effect size on the proximal phenotype is relatively small (for example, a 1-mg/dL change in plasma high-density lipoprotein cholesterol level), there is large locus heterogeneity, and, because of a large number of nongenetic factors, contributes to the phenotype.

In view of the shortcomings of clinical phenotyping and given the technical feasibility of whole exome or whole genome sequencing, one may infer that the era of a genome/exome-based approach to identify genetic determinants of the phenotype might not be far in the future. A desirable genotype-based approach will exploit the genome/exome data from thousands of individuals to prospectively link the genotype to phenotype through a comprehensive analysis and define the genetic architecture of human diseases and traits.

Whole genome sequencing is likely to become a commodity that could be readily available at a reasonable cost and be easily accommodated into the decision-making tree of health care of every individual. The challenging task will be to identify variants that are disease-causing or likely disease-causing and to develop strategies to prevent and attenuate the evolving phenotype (Figure 5). Likewise, various complementary studies, genetic and biological, would be necessary to discern the associated alleles from the true disease-causing variants. Moreover, a better understanding of various components of the genome, such as chromatin modification, functional conserved noncoding elements, transposons, large intergenic noncoding RNA, small noncoding RNA, and primary transcripts would be essential. An integrated approach that utilizes genetic, genomics, transcriptomics, proteomics, and metabolomics would be expected to facilitate
Identification and characterization of the mechanisms and involved in the pathogenesis of the phenotype.

**Sources of Funding**
Supported by NHLBI (R01-088498), NIA (R21AG038597), Burroughs Wellcome Award in Translational Research (1005907), and TexGen Fund from Greater Houston Community Foundation (to A.J.M.) and NHLBI R01 HL091771 and R01 HL090506 (to J.B.).

**Disclosures**
None.

**References**


Strategic Approaches to Unraveling Genetic Causes of Cardiovascular Diseases
A.J. Marian and John Belmont

Circ Res. 2011;108:1252-1269
doi: 10.1161/CIRCRESAHA.110.236067
Circulation Research is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2011 American Heart Association, Inc. All rights reserved.
Print ISSN: 0009-7330. Online ISSN: 1524-4571

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://circres.ahajournals.org/content/108/10/1252

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Circulation Research 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 Circulation Research is online at:
http://circres.ahajournals.org//subscriptions/