Moving Beyond Genome-Wide Association Studies

Nicole L. Glazer, PhD

The sequencing of the human genome and the subsequent completion of the human haplotype mapping (HapMap) project have led to unparalleled research into the structure and biology of the genome. These milestones held the promise of revolutionizing our understanding of the genetic basis for human disease, with the hope that population research would quickly lead to improved screening and diagnosis, along with targeted medical treatment. The advent of new genotyping technology and collaborative scientific efforts allowed for extensive scans across the genome to uncover genetic associations with various complex diseases and traits using genome-wide association (GWA) studies. At the end of 2010, the catalogue of published GWA studies compiled by the National Human Genome Research Institute (www.genome.gov/gwastudies) listed more than 700 published GWA studies and associations with nearly 4000 single-nucleotide polymorphisms (SNPs), providing new insights into the genetic architecture underlying disease. For the most part, however, the identified variants explain a relatively small proportion of the overall genetic variance of a given trait or disease, and the promise of a revolution in disease risk prediction and personalized genomic medicine has not yet been realized. Instead, what has occurred has been a revolution in biology rather than clinical medicine. Genetic studies have revealed a much greater complexity underlying the genomic basis for disease than scientists originally hypothesized. To maximize our understanding of these new biological mechanisms that may lead to advancements in the science of medicine, there is a need to follow up on GWA studies using multiple experimental methods and techniques with collaboration between different scientific disciplines.

Moving Toward Translation

Many significant associations discovered by GWA studies have not been within previously known candidate genes, thus identifying critical gaps in our knowledge and providing new clues to the pathophysiologic mechanisms underlying cardiovascular disease (CVD). However, when it comes to translation, it is necessary to not only localize the gene, but to follow up with characterization of the genes, identification of the functional variants, and interrogation of the functional mutations using experimental systems. Recent GWA studies conducted in >100,000 individuals identified 95 gene regions linked to plasma lipid levels. Although the SNPs most strongly associated with low-density lipoprotein cholesterol were in a noncoding region of chromosome 1p13, detailed functional genomics analysis revealed that a common mutation at the 1p13 locus creates a binding site for a transcription factor that binds in liver tissue and leads to enhanced expression of the SORT1 gene, some 120 kb away from the index SNP. Similar in-depth analysis of other GWA findings may uncover other novel regulators, but it is not always straightforward to tease apart how these gene influence trait levels or disease risk. In 2007, GWA studies identified common sequence variants on human chromosome 9p21 that confer an increased risk for coronary artery disease and myocardial infarction, but the gene region is devoid of protein-coding genes and the mechanism linking the region to coronary artery disease risk had remained unknown, preventing the development of molecular targeting strategies. Deletion of the corresponding stretch of DNA in mice has shown that this part of the chromosome regulates cardiac expression of two genes approximately 100 kb away. The downregulation of these genes, Cdkn2a/b, in a mouse results in excessive aortic smooth muscle cell proliferation, suggesting that the coronary artery disease susceptibility linked to 9p21 may occur through dysregulation of vascular cell proliferation. These types of multistage, multidisciplinary studies allow investigators to move from genomic localization to function to possible mechanism and illustrate the value and necessity of following up GWA study results. However, few studies have actually been able to localize the specific variant responsible for genetic associations.

Integration of Existing Data Sources

In addition to experimental studies, implementation of bioinformatics and utilization of existing GWA data sets and other genomic and biological databases to follow up on GWA discoveries probably will play an important role in the post-GWA phase. Implementation of bioinformatics and integration of existing data sources can allow for the validation of GWA finding, identification of causal variants, and development of the biological pathways underlying disease. The 1000 Genomes Project, with a goal of identifying the SNPs that are present at 1% or greater frequency, has allowed for follow-up of GWA studies through “fine mapping” by imputation. Liu et al conducted a meta-analysis of 20 studies to identify genes related to smoking. Subsequent interrogation of the smoking-related

Received January 5, 2011; accepted January 6, 2011.
From the Section of Preventive Medicine and Epidemiology, Department of Medicine, Boston University School of Medicine; and the Department of Epidemiology, Boston University School of Public Health, Boston, MA.
Correspondence to Nicole L. Glazer, PhD, Section of Preventive Medicine and Epidemiology, Boston University School of Medicine, 761 Harrison Ave, Boston, MA 02118. E-mail nlglazer@bu.edu

Circ Cardiovasc Genet is available at http://circgenetics.ahajournals.org

DOI: 10.1161/CIRCGENETICS.110.958785
locus using imputation with the 1000 genomes panel allowed for analysis of nearly all common SNPs in the region and offered a 5-fold increase in marker density over HapMap2, leading to the subsequent identification of the underlying causal variant, a SNP located within the promoter region of CHRNA5. In a GWA study of human height, bioinformatic and pathway-based approaches that used text-mining and gene set enrichment analysis found that the 180 loci identified were not random but were instead enriched for genes underlying skeletal growth and connected through biological pathways. To examine evidence for the potential involvement of specific genes, the authors aggregated their data from expression quantitative trait loci, pathway-based analyses, and existing human and mouse genetic data bases.

Despite the development of mega-GWA studies, including hundreds of thousands of individuals, increasing sample sizes for many cardiovascular traits is not possible because of limited phenotype assessment. At the same time, functional studies using “-omics” approaches like transcriptomics and proteomics are becoming essential for translating GWA findings, and the numbers of publicly available data bases are growing. Integration of these diverse data with existing GWA data sets, along with experimentation in model organisms, may allow for the identification of novel pathways and serve as a way to investigate the remaining missing heritability. The development and refinement of systems biology approaches will be crucial, particularly as these network data bases improve and multiply. Known metabolites and metabolite networks (metabolomics) are increasing, and the International Human Epigenome Consortium was recently launched, with an aim to map 1000 reference epigenomes within a decade. In some settings, existing GWA data sets may serve as replication data for pathways identified through systems biology analyses.

The Sequencing Rush

The modest proportion of heritable risk for CVD explained by variants identified through GWA studies has led to the question of whether a significant portion of the missing heritability is due to structural or rare variation. Technological advances and dropping costs are allowing for the rapid sequencing of both exomes and complete genomes. Data on rare alleles coming from the 1000 Genomes Project may also facilitate the investigation of rare alleles without the need for de novo sequencing. Both exome and whole-genome sequencing approaches have been successful in identifying rare variants in mendelian disorders through next-generation sequencing of small numbers of related and unrelated individuals. How and to what extent these strategies may be applicable to common, complex diseases such as CVD is unknown.

Complex CVD phenotypes will require sequencing relatively large numbers of individuals, and whole-genome sequencing on such a scale is still unaffordable. Targeted sequencing of all protein-coding regions in large numbers of people is more practicable, and the National Heart, Lung, and Blood Institute GO Exome Sequencing Project (https://esp.gs.washington.edu/drupal/) is currently underway, conducting whole-exome sequencing in well-phenotyped cardiovascular cohort studies with the aim of identifying rare and novel coding variants associated with CVD. Given that a large number of variants identified by GWA studies have been in noncoding regions, it will likely be necessary to complement sequencing of coding regions with targeted sequencing of select genomic regions and whole-genome sequencing in a limited number of individuals. The question of who are the optimal individuals to sequence is still unknown. To date, cardiovascular research has focused on sequencing those with rare genetic disorders, individuals at the extremes of a given phenotype, or those with diseases of unknown origin. The hope is that identification of genes involved in rare disorders or extreme phenotypes may provide targets for pharmacological or lifestyle interventions applicable to the general population.

Challenges in the Next Phase of Genomic Studies

To move from gene discovery to translation and advancements in medicine, there is a need to follow up on GWA discoveries using in-depth functional and mechanistic studies that cross multiple scientific disciplines. This will require collaboration on a large scale and the development of new analytic methods. Moving from genomic localization to function to mechanism is both time-consuming and expensive, and the identification of a single mechanism may not completely illuminate how a particular gene influences a CVD trait. Genetic loci identified through GWA studies may interact with each other or the environment in complex ways. An additional challenge will be how to make better use of existing data resources and how to deal with large amounts of data being currently produced. More and more institutions now have or will soon have large-scale sequencing facilities available to their researchers, leading to increasing data generation, identifying both real novel mutations along with false-positives. Availability of resources, including hardware, software, and personnel resources, is oftentimes the limitation to even routine analysis of genomics data. Given the significant number of loci that have been identified so far through GWA studies, follow-up will require a large amount of scientific investment, but one that is necessary in moving from gene discovery to translation.

Disclosures

None.

References

O, Koteliansky V, Fitzgerald K, Krauss RM, Cowan CA, Kathiresan S,
KV, Li X, Li H, Kuperwasser N, Ruda VM, Pirruccello JP, Muchmore B,

KV, Li X, Li H, Kuperswasser N, Ruda VM, Pirruezzo JP, Muchmore B,
Prokunina-Olsson L, Hall JL, Schadt EE, Morales CR, Kuehnert S, Phillips
MC, Wong J, Cantley W, Ruda VM, Kayser M, Kuchel DA, Kato N, Kurokawa K,
Kotelnikov V, Kusama K, Lintoglou E, Kotsis G, Kovanagi A, Kriek M,
Kromann H, Kromann J, Kryukov GV, Kwan M, Kruglyak L, Kugelberg E,
Kullikova I, Kullikov N, Kullikov A, Khurana HS, Hoffmann K, Mosley TH Jr,
Hoffman W, Helgason AM, Helgason T, Helgason J, Helgason T, Helgason S,
Helgason B, Helgason B, Helgason A, Hemmingson M, Hemmingson M,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J, Hemmingson J,
Moving Beyond Genome-Wide Association Studies
Nicole L. Glazer

Circ Cardiovasc Genet. 2011;4:91-93
doi: 10.1161/CIRCGENETICS.110.958785
Circulation: Cardiovascular Genetics is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2011 American Heart Association, Inc. All rights reserved.
Print ISSN: 1942-325X. Online ISSN: 1942-3268

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://circgenetics.ahajournals.org/content/4/1/91

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