ANNUAL REVIEWS **Further**

Click here for quick links to Annual Reviews content online, including:

- Other articles in this volume
- Top cited articles
- Top downloaded articles
- Our comprehensive search

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

The HapMap and Genome-Wide Association Studies in Diagnosis and Therapy*

Teri A. Manolio and Francis S. Collins

National Human Genome Research Institute, Bethesda, Maryland 20892;
email: manolio@nih.gov

Annu. Rev. Med. 2009. 60:443–56

The *Annual Review of Medicine* is online at med.annualreviews.org

This article's doi:
10.1146/annurev.med.60.061907.093117

Copyright © 2009 by Annual Reviews.
All rights reserved

0066-4219/09/0218-0443$20.00

*The U.S. Government has the right to retain a nonexclusive, royalty-free license in and to any copyright covering this paper.

Key Words

complex diseases, genetic association, genomic variation

Abstract

The International HapMap Project produced a genome-wide database of human genetic variation for use in genetic association studies of common diseases. The initial output of these studies has been overwhelming, with over 150 risk loci identified in studies of more than 60 common diseases and traits. These associations have suggested previously unsuspected etiologic pathways for common diseases that will be of use in identifying new therapeutic targets and developing targeted interventions based on genetically defined risk. Here we examine the development and application of the HapMap to genome-wide association (GWA) studies; present and future technologies for GWA research; current major efforts in GWA studies; successes and limitations of the GWA approach in identifying polymorphisms related to complex diseases; data release and privacy polices; use of these findings by clinicians, the public, and academic physicians; and sources of ongoing authoritative information on this rapidly evolving field.

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

GWA: genome-wide association

Single nucleotide polymorphism (SNP): site within the genome that differs by a single nucleotide base across different individuals

Polymorphism: a form of genetic variation in which each allele occurs in at least 1% of the population

Tag SNP: representative SNP in a region of the genome with high linkage disequilibrium to other variants

Linkage disequilibrium (LD): association of alleles at two or more sites on the same chromosome that are inherited together more often than expected by chance

Haplotype: a combination of alleles at multiple linked sites on a single chromosome that are transmitted together

THE HAPMAP: BUILDING THE FOUNDATION FOR GENOME-WIDE ASSOCIATION STUDIES

The International HapMap Project was designed to create a public, genome-wide database of patterns of common human sequence variation to guide genetic studies of human health and disease, including genome-wide association (GWA) studies (1, 2). Identifying genetic influences on complex diseases would be quite difficult if the risk-associated allelic variants at a particular disease-causing locus were very rare, so that for a disease to be common there would be many different causative alleles. The HapMap was instead designed to facilitate identification of commonly occurring disease-causing variants based upon the "common disease, common variant" hypothesis (3). This hypothesis suggests that at least some of the genetic influences on many common diseases are attributable to a limited number of common allelic variants that are present in more than 5% of the population.

GWA studies attempt to identify these common disease-causing variants by using high-throughput genotyping technologies to assay hundreds of thousands of common single nucleotide polymorphisms (SNPs) and relate them to clinical conditions and measurable traits. Because of the strong associations among SNPs in most chromosomal regions, only a few carefully chosen SNPs need to be typed in each region to predict the likely alleles at the rest of the SNPs in that region. Selecting the best tag SNPs requires precise mapping of the patterns of linkage disequilibrium (LD) among SNPs, which differ somewhat across ancestral groups. The need for precise LD maps to facilitate genetic association studies was the stimulus for developing the human haplotype map (2, 4).

The International HapMap Project was a consortium of researchers in Canada, China, Japan, Nigeria, the United Kingdom, and the United States, organized to produce a human haplotype map by genotyping 270 samples from four populations with geographically diverse ancestry (1, 2). These samples included 30 mother–father–adult child trios from the Yoruba in Ibadan, Nigeria; 30 trios from the CEPH (Centre d'Etude du Polymorphisme Humain) collection of Utah residents of Northern and Western European ancestry; 45 unrelated Han Chinese individuals in Beijing, China; and 45 unrelated Japanese individuals in Tokyo, Japan. Approximately 1 million SNPs were genotyped and their LD patterns characterized in Phase I of the project. A description was published in 2005 (1), but the data were available long before this and were central to several early genomic discoveries in complex diseases (5, 6). The Phase II HapMap of more than 3 million SNPs was published in 2007 (7).

Subsequent research has shown that tag SNPs chosen using the HapMap are generally applicable across other populations, but there are some limitations, particularly for rarer SNPs and for populations with substantial proportions of recent African ancestry (8). To allow better choice of tag SNPs and more detailed analyses for diverse populations, additional samples were collected from the same four initial HapMap populations and from seven additional populations: Luhya in Webuye, Kenya; Maasai in Kinyawa, Kenya; Tuscans in Italy; Gujarati Indians in Houston, Texas; Chinese in metropolitan Denver, Colorado; persons of Mexican ancestry in Los Angeles, California; and persons of African ancestry in the Southwestern United States (9). These 1301 extended HapMap samples are now available from the Coriell Institute and have been genotyped on the Affymetrix 6.0 platform and the Illumina 1 million SNP chip. Genome-wide sequencing of these samples to develop a comprehensive catalog of rarer variants will begin soon as part of the international 1000 Genomes Project (**http://www.1000genomes.org**).

GENOME-WIDE ASSOCIATION TECHNOLOGIES, PRESENT AND FUTURE

GWA studies have been defined by the National Institutes of Health (NIH) as any studies

of common genetic variation across the entire human genome designed to identify genetic associations with observable traits (10). Implicit in this definition is that sufficient numbers of SNPs are typed to capture the vast majority of common variations (as noted above, these are alleles with a frequency of at least 5% in a population) throughout the entire genome. Such studies typically involve hundreds of thousands of SNPs and are not limited to known genes or regulatory regions. Instead, they assess genetic variation genome-wide in an almost "agnostic" fashion, unconstrained by current imperfect understanding of genome structure and function (11).

Technologies for high-throughput assays of thousands, and then tens and hundreds of thousands, of SNPs developed in parallel with the progress of the HapMap, as it became clear that denser maps could effectively capture the majority of human genetic variation (12, 13). These advances have made possible the dense genotyping needed to characterize the SNP variation within an individual, at a sufficiently low cost to allow the large sample sizes needed for comparisons of persons with and without disease. As genotyping platforms expand to include ever more tag SNPs, they capture increasingly larger proportions of the variation in any population, so that even samples of recent African ancestry, characterized by greater variation and shorter stretches of LD (14), have most of the genome covered at high r^2 (7).

Current-generation high-throughput genotyping platforms are extraordinarily efficient at genotyping SNPs but are less effective at genotyping structural variants, such as insertions, deletions, inversions, and copy number variants (CNVs). These variants are common in the human genome, though not as common as SNPs (15). The HapMap was not designed to capture these variants, although it can be used indirectly to do so, particularly for deletions that are in strong LD with SNPs (16). CNVs, in which stretches of genomic sequence roughly 1 kb to 3 Mb in size are deleted or duplicated in varying numbers, have gained increasing attention because of their apparent ubiquity and potential dosage effect on gene expression (17).

A critical question related to methods for typing CNVs is whether they usually arise from a single originating event and then are propagated with diminishing degrees of LD on a single haplotype background, or instead are frequently regenerated on varying haplotype backgrounds (18). The former situation would be conducive to tagging and indirect interrogation with HapMap-based genotyping platforms, while the latter would likely require direct interrogation or genomic sequencing for reliable association studies. Expansions and refinements of current genotyping platforms are increasingly focused on capturing CNVs adequately, and some success has already been achieved (19). Array and sequencing methods are also being used to type structural variants, using the HapMap samples for development and cross-validation of the methods (20, 21).

The identification of rare, potentially causal variants that are poorly tagged by existing genotyping platforms will require sequencing DNA from large numbers of people for the genomic regions showing strong associations with complex traits (22). The 1000 Genomes Project plans to produce modest sequence coverage (an average of four sequencing reads at any place in the genome) of ~1500 individuals that will extend the catalog of human genetic variation to variants present in 1%–5% of the population (10). It will thus limit the follow-up sequencing needed for investigating specific association findings to the search for very rare variants. Fine-mapping of candidate regions with common and rare SNPs optimally chosen, based on HapMap data, to maximize the regional genomic variation captured while minimizing costs, will refine association signals and narrow the list of possible functional variants.

r^2: Linkage disequilibrium coefficient representing the proportion of observations in which two specific pairs of alleles occur together

Copy number variant (CNV): a DNA sequence of hundreds to thousands of base pairs that occurs a variable number of times across individuals

CURRENT MAJOR EFFORTS IN GENOME-WIDE ASSOCIATION STUDIES

The first association study generally considered to be truly genome-wide was published in

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

Nonsynonymous SNP: a SNP for which each allele encodes a different amino acid in the protein sequence

March 2005 (23), and by August 2008 >170 such publications had identified >150 genetic loci associated with >60 complex diseases and traits (9, 24). Although many such efforts have been and will continue to be undertaken individually, or as part of single large-scale studies such as the deCODE database (25) or the National Heart, Lung, and Blood Institute's Framingham Study (26), the value of collaborative efforts across studies and even across diseases is increasingly being recognized.

A series of coordinated GWA publications in early 2007 in prostate cancer, breast cancer, and myocardial infarction demonstrated the value of assessing association reports in multiple studies simultaneously (9). The joint publication of individual and combined associations with type 2 diabetes in three collaborating studies definitively showed the importance of combining individual-level genotype and phenotype data in >30,000 subjects to identify associations across several studies that no single study could reliably identify on its own (27–29). This approach was subsequently expanded by the addition of seven more diabetes studies in the Diabetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, with an effective sample size of >50,000 (30). Similarly large efforts focused on a single phenotype or closely related phenotypes in tens of thousands of subjects have yielded variants related to obesity (31), lipids (32), height (33), and other traits.

A more challenging and somewhat more controversial approach has been to combine individual-level data from cases with several related or even unrelated conditions and compare them to a common control group, in an effort to expand sample size and increase study power for each condition. The success of this method was demonstrated by the landmark Wellcome Trust Case Control Consortium (WTCCC) study of 2000 cases of each of seven common diseases and 3000 shared controls (34). This study provided many fundamental methodologic advances, including demonstration of the robustness of a single control group, the value of using cases of some diseases as controls for others, the greater power provided by increased sample size (numbers of subjects) rather than increased genomic coverage (numbers of SNPs), the critical need for manual review of automated genotyping calls, and the reliability of imputed genotypes for SNPs that were not actually typed by the genotyping platform. This approach of common controls and combined case groups used as controls was also employed by the WTCCC in its smaller study of 14,500 nonsynonymous SNPs in four autoimmune diseases, and in dense genome-wide genotyping of African cases of tuberculosis and malaria (35, 36). The Wellcome Trust recently announced plans to conduct genome-wide genotyping in 120,000 additional people to identify variants related to 25 diseases and traits using the same approach (37).

Other collaborative studies of multiple diseases have focused less on combining genotype-phenotype associations than on sharing methods for genotyping quality control, data analysis, imputation, and data distribution. Experience gained from early quality-control efforts in programs such as the Genetic Association Information Network (GAIN) of six complex diseases has been of great value in speeding the completion and analysis of genotyping in later studies (19). Several other collaborative programs are currently in the pipeline (**Table 1**).

SUCCESSES IN IDENTIFYING VARIANTS RELATED TO COMPLEX DISEASES

The first notable success of the GWA method came in March 2005, with the identification of a variant in the gene for complement factor H (CFH) associated with age-related macular degeneration (23). Two additional GWA studies were published within that year, of Parkinson's disease and obesity (38, 39), but efforts at replicating these findings have produced inconsistent results (40, 41). In 2006, strong, robust associations with electrocardiographic QT interval prolongation (42), neovascular macular degeneration (43), and inflammatory bowel

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

Table 1 Collaborative genome-wide association studies (adapted from Reference 9)

Study name	Genetic Association Information Network (GAIN)	Genes, Environment, and Health Initiative (GEI)	SNP Typing for Association with Multiple Phenotypes from Existing Epidemiologic Data (STAMPEED)	Cancer Genetic Markers of Susceptibility (CGEMS)	Psychiatric Genomewide Association Study Consortium (PGC)
URL	http://www.fnih.org/GAIN2/home_new.shtml	http://www.gei.nih.gov/	http://public.nhlbi.nih.gov/GeneticsGenomics/home/stampeed.aspx	http://cgems.cancer.gov/	http://sullivanlab.unc.edu/pgc/index.html
Traits or diseases studied	attention deficit/hyperactivity disorder	type 2 diabetes	early-onset myocardial infarction	prostate cancer	autism
	major depressive disorder	maternal metabolism and birth weight	asthma	breast cancer	attention deficit/hyperactivity disorder
	bipolar I disorder	preterm birth	platelet phenotypes	pancreatic cancer	bipolar disorder
	schizophrenia	oral clefts	coronary heart disease and other heart, lung, and blood disorders	lung cancer	major depressive disorder
	type 1 diabetic nephropathy	dental caries	childhood respiratory outcomes	bladder cancer	schizophrenia
	psoriasis	coronary disease	hematopoietic cell transplant outcome	renal cancer	
		lung cancer	arteriosclerosis in hypertensives		
		addiction	asthma and lung function		
			cardiovascular risk factors		
			atherosclerosis pathway genes		
			cardiovascular events		
			early coronary artery disease		
			phenotypic variability in sickle-cell anemia		
			longevity to age 100		

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

disease (44) were identified and have since been the subjects of a substantial body of follow-up research to determine gene function and population impact.

The pace of genomic discovery increased dramatically in 2007, following the increased availability of high-density genotyping platforms and experience in interpreting the results. Simultaneous publication of coordinated efforts in multiple diseases, and of the WTCCC study, have been described above. Rapid progress has continued into 2008 with identification of >150 loci for >60 common diseases and traits (**Figure 1**). Indeed, as Hunter & Kraft have noted, "There have been few, if any, similar bursts of discovery in the history of medical research" (45).

Unique aspects of the GWA method have made these discoveries possible. For example, GWA studies allow the investigator to narrow an association region to a 10–100 kilobase length of DNA, in contrast to the 5–10 megabases usually detected in familial linkage studies. Because GWA regions typically contain only a few genes, rather than the dozens or hundreds implicated in linkage regions, potentially causative variants can be examined much more rapidly and in greater depth. As noted above, systematic interrogation of the entire genome frees the investigator from reliance on inaccurate prior hypotheses based on incomplete understanding of disease pathogenesis and genome structure and function. The critical importance of this is illustrated by the fact that many of the associations identified to date, such as CFH in macular degeneration (23) and TCF7L2 in type 2 diabetes (6, 46), have been surprising—the genes were not previously suspected of being related to the disease. Some, such as the strong associations of prostate cancer with SNPs in the 8q24 region (47) and Crohn's disease with the 5p13 region (34), have been in genomic regions containing no known genes at all. And because current genotyping assays capture the vast majority of human variation genome-wide, rather than being focused on particular regions or pathways, once a GWA scan is completed it can be applied to any condition or trait measured in that same individual and consistent with his or her informed consent.

Several of these discoveries have suggested etiologic pathways not previously implicated in these diseases, such as the autophagy pathway in inflammatory bowel disease (48), the complement pathway in macular degeneration (23), and the HLA-C locus in control of viral load in HIV infection (49). Of considerable interest in determining pathophysiology have been variants or regions implicated in multiple diseases, such as the 8q24 region in prostate, breast, and colorectal cancers and the *PTPN2* gene in type 1 diabetes and Crohn's disease (9).

LIMITATIONS OF THE GENOME-WIDE ASSOCIATION METHOD

Important limitations of GWA studies should also be kept in mind. One is their enormous potential for generating false-positive or spurious associations. Because they test hundreds of thousands of statistical hypotheses—one for each allele or genotype assessed—GWA studies have enormous potential for generating false-positive results due to chance alone. At the usual $p < 0.05$ level of significance, an association study of one million SNPs will show 50,000 SNPs to be "associated" with disease, almost all spuriously. One response to this problem is to reduce the false-positive rate by applying the Bonferroni correction, in which the conventional p-value is divided by the number of tests performed (45). A one-million-SNP survey would thus use a threshold of $p < 0.05/10^6$, or 5×10^{-8}, to identify associations unlikely to have occurred by chance. This correction has been criticized as overly conservative, but it remains the most commonly used approach to date (50). Another cause of the false-positive associations to which GWA studies are prone is population stratification. Allele frequencies vary between population subgroups, such as those defined by ethnicity or geographic origin,

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

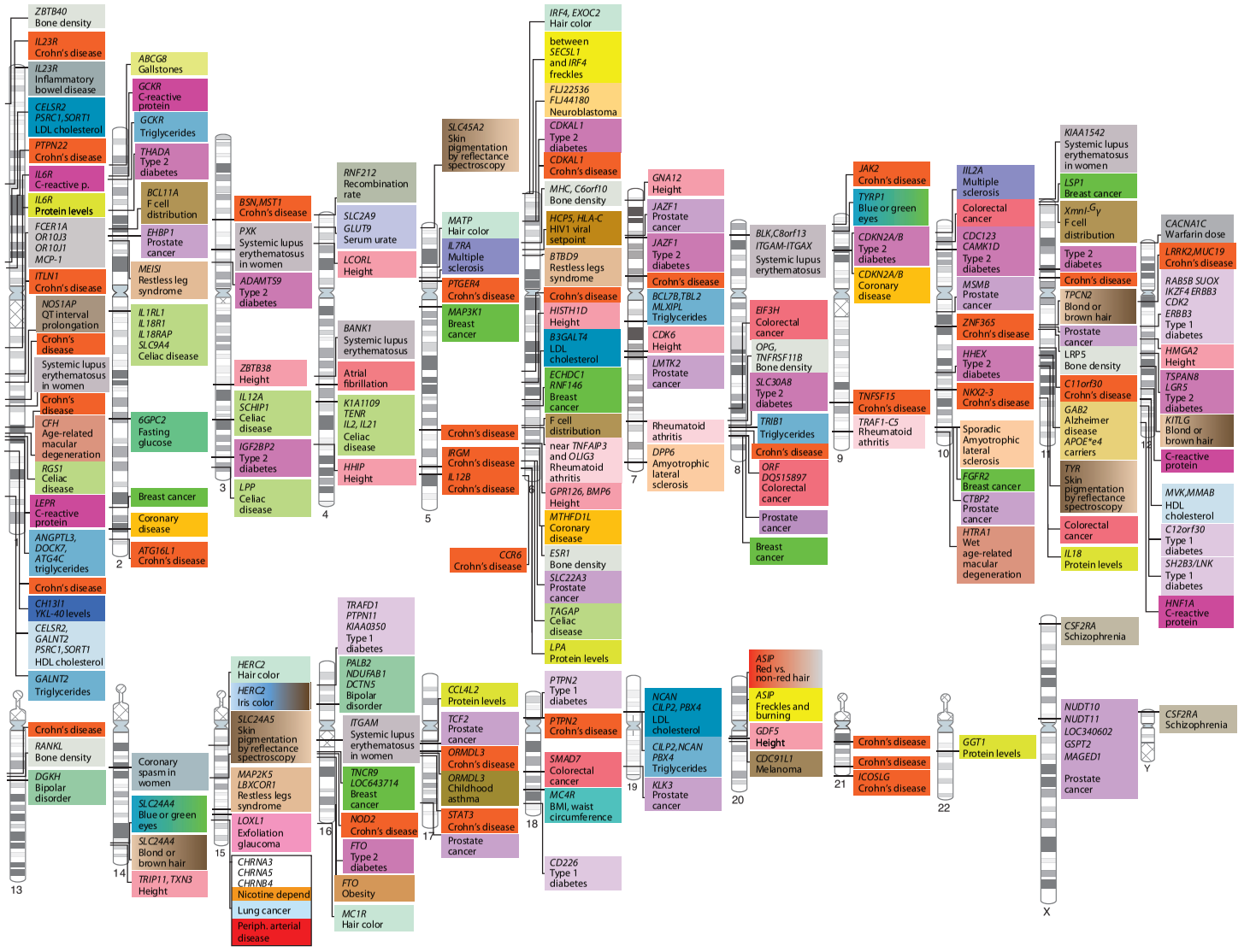

Figure 1

SNP-trait associations detected in GWA studies. Associations significant at $p < 9.9 \times 10^{-7}$ and reported through June 2008 are shown according to chromosomal location and involved or nearby gene, if any. Colored boxes indicate similar diseases or traits. Adapted from Reference 9 with permission.

and these subgroups in turn differ in their risk for disease. GWA studies may then falsely identify the subgroup-associated genes as related to disease (50). Genotyping error is another important cause of spurious associations that must be carefully sought and corrected.

dbGaP: Database of Genotype and Phenotype

These problems have also plagued candidate-gene association studies, where systematic reviews showed that the vast majority of initial associations could not be replicated (51). This experience has led to calls for all genetic association reports to include documented replication of findings as a prerequisite for publication. Consensus guidelines for replication in any GWA study, and, crucially, for complete description of the initial study so that replication is possible, have been developed (52).

Another limitation of GWA studies is their lack of power for identifying associations with rare sequence variants, since these are poorly represented on current genotyping platforms, as are structural variants. The often limited information available on environmental exposures and other nongenetic risk factors in GWA studies will make it difficult to identify gene-environment interactions, or modification of gene-disease associations in the presence of environmental factors. Most variants identified to date are of relatively modest effect size, conferring less than a twofold increase in disease risk and necessitating large sample sizes to detect their effect (9, 34). Although the importance of risk factors with small effect is debated, the potential for purifying selection to eliminate risk variants of large effect is unique to genetic studies and will tend to keep effect sizes small for common variants (53). Modest associations can point the way to important therapeutic avenues, and, when considered in combination, may identify persons at substantially increased risk (28). Such information can be particularly important, even in the absence of specific pharmaceutical agents targeted to such individuals, for more aggressive efforts to reduce known risk factors that are modifiable, such as obesity in prediabetes and smoking in age-related macular degeneration (9).

DATA RELEASE AND PARTICIPANT PRIVACY

GWA studies produce massive data sets, often representing substantial investments of public funds and providing unparalleled opportunities for research into complex diseases. Recognizing the research potential of these data sets, and following an extended period of public comment, NIH released its Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies, recommending widespread and responsible release of GWA data to the scientific community through the Database of Genotype and Phenotype (dbGaP) of the National Center for Biotechnology Information (10, 54). Study descriptions and protocols are available in the open-access portion of the dbGaP website; individual-level data are provided through a controlled-access process consistent with participants' informed consent. This commitment to rapid data release builds on the now well-established ethic in genomic community research projects of maximizing data access. Other GWA data access sites include the Cancer Genetic Markers of Susceptibility (CGEMS) data portal (55) and the European Genotype Archive (EGA) (56). Policies for data release have been developed collaboratively among these projects and are quite similar. Published GWA studies and major findings are also catalogued by the National Human Genome Research Institute (NHGRI) (24), and GWA literature citations are available through the Centers for Disease Control and Prevention (57).

The extensive genotype and phenotype information deposited in dbGaP raises important questions about possible risks to confidentiality of individual participants in broad data-sharing models. NIH policies were thus developed with deliberate attention to participant protections, both in the process of data submission from the original studies and in the processes of data access and use by outside investigators. A key aspect of the protections provided in dbGaP is removal of potentially identifying information

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

prior to data submission, using criteria very similar to those described within the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule (10).

Substantial participant protections are also applied at the dbGaP data-user level through a process managed by a Data Access Committee (DAC) composed of senior NIH staff. Investigators interested in obtaining controlled-access dbGaP data submit a Data Access Request, cosigned by their institution, constituting their agreement to abide by the principles and practices detailed in the NIH GWA study policy. These include keeping the data secure; using them only for the approved research purposes; acknowledging NIH policies on publications and intellectual property (IP); and submitting periodic reports on data use. Data users also agree not to distribute individual-level data in any form to any third parties (other than their own research staff who have agreed to the terms of access), nor to attempt to identify individual study participants.

Recognizing the unprecedented pace of scientific progress in this field, NIH has designed its policies on data sharing in GWA studies to adapt to rapid technical advances. For example, data summaries and other group-level data such as allele frequencies and association statistics were initially provided in the open-access portion of dbGaP, in the belief that grouped data carried no threats to individual privacy. An innovative analysis for resolving the presence of an individual's DNA in a mix of DNA (as from a mass disaster) subsequently showed that an individual could be determined to have contributed to grouped allele frequency data with high reliability if one had data on that individual's genotypes at hundreds of thousands of SNPs (58). NIH responded swiftly to remove these data sets from open access and place them behind the controlled-access process, and to notify other major data providers as well, who took similar actions (59). Data access policies will continue to evolve with ongoing scientific advances in the field, to ensure that state-of-the-art data can be distributed and used in the most responsible and productive manner.

Ensuring confidentiality and privacy is vital for databases containing individual-level genotype or phenotype information. Important concerns about the potential for persons carrying risk-associated variants (i.e., essentially everyone) becoming the object of discrimination by employers or insurers must be addressed. Application of GWA findings and other genomic research will be greatly facilitated by the formal legal protection against discrimination based on genetic information provided by the Genetic Information Nondiscrimination Act (GINA) that was signed into law in May 2008. When it goes into effect in 2009, GINA will protect against discrimination by health insurers and employers on the basis of genetic information. It is particularly important for GWA studies because of the breadth of information obtained in such studies; almost certainly every individual will carry at least one risk allele for at least one common disease. Protections under GINA will not only shield study subjects from the risk of genetic discrimination due to their participation in research, but more importantly will ensure that clinicians can order genetic tests identified from GWA studies to make more effective treatment decisions and can place this information in patients' records without risk to patients or their families.

USE OF FINDINGS BY PHYSICIANS AND THE PUBLIC

Although GWA discovery studies provide valuable clues to genomic function and pathophysiologic mechanisms, they are only a first step in identification of disease genes and are many steps removed from actual clinical application. Nonetheless, they tend to receive considerable media attention and have the potential for generating queries from patients about whether to get tested for the "new gene for Disease X" based on the latest report (50). As noted above, many SNPs identified from such studies, as well as the genes or regions containing them, are currently of unknown function. In addition, SNPs from GWA studies in complex diseases (unlike many Mendelian disorders) do not

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

predict unequivocally who will develop disease and who will remain free of it. Instead, individuals carrying a particular risk genotype implicated in a GWA study have a greater risk (and sometimes only a modestly greater risk) of developing a complex disease than those who do not.

The distinction between disease prediction and disease susceptibility is important because for many common variants, a substantial number of persons who do not carry the at-risk genotype may develop disease anyway owing to environmental or other factors. Indeed, for common diseases such as hypertension or diabetes, environmental or lifestyle factors may play such a strong role relative to genetics that many individuals with the at-risk genotype will develop disease for reasons that are probably unrelated to genotype, and others with the at-risk genotype may remain healthy in the absence of other important environmental exposures (60). Identifying subgroups of individuals in whom SNP-outcome associations differ according to the presence or absence of other SNPs or environmental factors might eventually be of considerable clinical use, particularly for environmental factors that can be modified.

The consensus at present is that GWA findings provide important clues to disease etiology and pathways to treatment, but current information is far too preliminary to recommend their use in prevention or treatment recommendations. Use of GWA findings in screening for disease risk, though beginning to be marketed commercially, is problematic. Although getting the latest "gene test" may be alluring, evidence is needed that such screening adds information to known risk factors (such as age, smoking, obesity, and family history), that effective interventions are available, that improved outcomes justify the associated costs, and that obtaining this information does not have serious adverse consequences for patients and their families.

Given the availability of many genetic tests to anyone willing to pay for them, however, clinicians are soon likely to face anxious patients equipped with genotype information showing them to be at risk for multiple diseases. This may provide a "teachable" moment for encouraging patients to apply known preventive strategies against the conditions for which they are at increased risk. Such encounters also provide critical opportunities to discourage complacency in preventive strategies for which genotyping information suggests a patient is not at increased risk. This is because so little is known about genetic influences on complex diseases and because variants identified to date typically explain so small a proportion of population risk. It may be useful to point out to patients considering purchasing these tests that obtaining a family history is often simpler and almost always cheaper. A positive family history typically confers a three- to fourfold increase in the risk of many diseases and is extremely useful in identifying persons to target for more intensive screening (61).

UTILITY OF FINDINGS FOR RESEARCH

Perhaps the greatest initial utility of GWA findings is in the clues they provide for disease etiology, therapeutic targets, and gene function. As noted above, several of these discoveries have suggested etiologic pathways and therapeutic opportunities not previously implicated in the complex diseases with which they are associated, such as the autophagy pathway in inflammatory bowel disease, the complement pathway in macular degeneration, and the HLA-C locus in control of viral load in HIV infection (9). Intriguing potential genetic connections between diseases previously believed to be unrelated—such as the finding that risks of type 2 diabetes, coronary disease, and familial melanoma are all associated with variants near *CDKNA2A/B*, or that risks of Crohn's disease and type 1 diabetes are related to variants near *PTPN2*—suggest new avenues of research in identifying other similarities in etiology, progression, or treatment of these conditions. The not infrequent occurrence of associations in "gene deserts" far from any known genes invites the question of whether studies of disease

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

pathogenesis have been too focused on coding regions of the genome and have missed other important structural and functional clues to genomic regulation.

Research to pursue initial GWA discoveries will include replication studies in the same phenotypes and populations, to ensure the robustness of the findings, and in similar but not identical phenotypes and populations, to extend the findings and increase understanding of their mechanisms and importance (52). Investigation of disease subtypes, such as estrogen receptor–positive versus –negative breast cancer, or young-onset or severely progressive forms of prostate cancer or diabetes, may be of great value in identifying which subgroups of alleles confer the highest risk and which subgroups of persons carry those alleles. Functional studies of highly replicated variants, in experimental models such as knockdown and overexpression studies (9) and in relationship to gene expression, as recently demonstrated for asthma-associated variants in *ORMDL3* (62), will help to determine the mechanisms of gene function and how they are perturbed in disease, providing insights into possible preventive or therapeutic strategies.

AUTHORITATIVE SOURCES OF INFORMATION

The HapMap continues to evolve, with new SNPs being identified and LD patterns defined in both the original and newer HapMap populations. The primary portal to HapMap genotype data, as well as publications, tutorials, and other relevant resources, is the International HapMap Project website at **http://www.hapmap.org**. An up-to-date catalog of GWA studies is provided by the NHGRI's Office of Population Genomics at **http://www.genome.gov/GWAstudies**. This site lists all published studies attempting to assay 100,000 SNPs or more, noting the trait under investigation, the top new associations identified, their genomic region and nearby genes, *p*-values, odds ratios, and links to the PubMed citations. Study descriptions, protocols, and association findings are available for many NIH-supported GWA studies in dbGaP at **http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap**, and individual-level data may be requested for download through the controlled-access portion of that site. Those seeking additional information on specific genes related to complex diseases should consult Online Mendelian Inheritance in Man (OMIM) at **http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM**, the definitive catalog of human genes and genetic disorders. Findings from GWA studies are added to genes described in OMIM on a regular basis. More relevant to clinicians and patients may be the website and materials produced by the National Coalition for Health Professional Education in Genetics, a coalition of health professional organizations whose purpose is to promote health professional education and access to information about advances in human genetics, at **http://www.nchpeg.org**.

SUMMARY

The genome-wide database of human genetic variation produced by the International HapMap Project has provided a radically new approach for searching for genetic variants associated with complex diseases. The overwhelming success of these studies has led to surprising new insights into disease pathophysiology and therapeutic approaches, as well as new questions about genomic structure and function (and its interaction with genomic variation and environmental factors) in disease causation.

GWA studies represent a powerful new tool for identifying genetic variants related to complex diseases, but they also have important limitations, including their potential for false-positive results and lack of sensitivity to detect rare variants. Their primary uses for the foreseeable future are likely to be in the investigation of biologic pathways of disease causation and normal health and development. Clinical application of these findings will require firm evidence that testing for them adds information to known risk factors, that effective interventions are available, that improved

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

outcomes justify the associated costs, and that obtaining this information does not have serious adverse consequences for patients and their families. Although most GWA findings are clearly several steps removed from mainstream clinical use at present, functional investigation and experimental application of these findings are expected to produce new advances in the prevention and treatment of common diseases.

FUTURE ISSUES

1. Defining the functional properties of genomic variants identified through GWA studies, including effects on gene expression, protein structure, and protein function.
2. Identifying copy number variants that may be contributing significantly to common disease risk but are scored inconsistently by current technologies.
3. Identifying rarer sequence variants that may be causative or additive in the disease associations identified in GWA studies.
4. Determining the population prevalence and risk associated with putative causal variants in unbiased and diverse population samples.
5. Estimating the increment in risk over established risk factors provided by GWA-defined variants.
6. Using information from GWA studies to identify new targets for therapeutic intervention.

DISCLOSURE STATEMENT

The authors are not aware of any factors that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

1. International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* 437:1299–1320
2. International HapMap Consortium. 2003. The International HapMap Project. *Nature* 426:789–94
3. Collins FS, Guyer MS, Chakravarti A. 1997. Variations on a theme: cataloging human DNA sequence variation. *Science* 278:1580–81
4. Eberle MA, Ng PC, Kuhn K, et al. 2007. Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet.* 3:1827–37
5. Benusiglio PR, Lesueur F, Luccarini C, et al. 2005. Common variation in EMSY and risk of breast and ovarian cancer: a case-control study using HapMap tagging SNPs. *BMC Cancer* 5:81
6. Grant SF, Thorleifsson G, Reynisdottir I, et al. 2006. Variant of transcription factor 7–like 2 (TCF7L2) gene confers risk of type 2 diabetes. *Nat. Genet.* 38:320–23
7. International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–61
8. deBakker PI, Burtt NP, Graham RR, et al. 2006. Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.* 38:1298–1303
9. Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* 118:1590–605
10. 2007. Policy for sharing of data obtained in NIH supported or conducted genome-wide association studies (GWAS). Federal Register 8/30/07. **http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html**, accessed 9/12/2008

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

11. Carlson CS. 2006. Agnosticism and equity in genome-wide association studies. *Nat. Genet.* 38:605–6
12. Wang DG, Fan JB, Siao CJ, et al. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280:1077–82
13. Matsuzaki H, Dong S, Loi H, et al. 2004. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods* 1:109–11
14. Gabriel SB, Schaffner SF, Nguyen H, et al. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–29
15. Tuzun E, Sharp AJ, Bailey JA, et al. 2005. Fine-scale structural variation of the human genome. *Nat. Genet.* 37:727–32
16. Komura D, Shen F, Ishikawa S, et al. 2006. Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* 16:1575–84
17. Stranger BE, Forrest MS, Dunning M, et al. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* 315:848–53
18. Estivill X, Cox NJ, Chanock SJ, et al. 2008. SNPs meet CNVs in genome-wide association studies: HGV2007 meeting report. *PLoS Genet.* 4:e1000068
19. Manolio TA, Rodriguez LL, Brooks L, et al. 2007. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nat. Genet.* 39:1045–51
20. Estivill X, Armengol L. 2007. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. *PLoS Genet.* 3:1787–99
21. Kidd JM, Cooper GM, Donahue WF, et al. 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature* 453:56–64
22. Frayling TM, McCarthy MI. 2007. Genetic studies of diabetes following the advent of the genome-wide association study: Where do we go from here? *Diabetologia* 50:2229–33
23. Klein RJ, Zeiss C, Chew EY, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* 308:385–89
24. National Human Genome Research Institute. *A catalog of genome-wide association studies.* **http://www.genome.gov/GWAstudies/**, accessed 4/28/08
25. Gulcher J, Kong A, Stefansson K. 2001. The genealogic approach to human genetics of disease. *Cancer J.* 7:61–68
26. Cupples LA, Arruda HT, Benjamin EJ, et al. 2007. The Framingham Heart Study 100K SNP genome-wide association study resources: overview of 17 phenotype working group reports. *BMC Med. Genet.* 8(Suppl. 1):S1
27. Saxena R, Voight BF, Lyssenko V, et al. 2007. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 316:1331–36
28. Scott LJ, Mohlke KL, Bonnycastle LL, et al. 2007. A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 316:1341–45
29. Zeggini E, Weedon MN, Lindgren CM, et al. 2007. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 316:1336–41
30. Zeggini E, Scott LJ, Saxena R, et al. 2008. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* 40:638–45
31. Frayling TM, Timpson NJ, Weedon MN, et al. 2007. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* 316:889–94
32. Kathiresan S, Melander O, Guiducci C, et al. 2008. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat. Genet.* 40:189–97
33. Weedon MN, Lango H, Lindgren CM, et al. 2008. Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.* 40:575–83
34. Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–78
35. Wellcome Trust Case Control Consortium; Australo-Anglo-American Spondylitis Consortium (TASC), Burton PR, Clayton DG, Cardon LR, et al. 2007. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* 39:1329–37
36. Wellcome Trust Case Control Consortium. *Overview*. **http://www.wtccc.org.uk/info/overview.shtml**, accessed 4/29/08

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

37. Wellcome Trust. 2008. *Largest ever study of genetics of common diseases just got bigger*. News release. **http://www.wellcome.ac.uk/News/Media-office/Press-releases/2008/WTD039438.htm**, accessed 4/29/08
38. Maraganore DM, de Andrade M, Lesnick TG, et al. 2005. High-resolution whole-genome association study of Parkinson disease. *Am. J. Hum. Genet.* 77:685–93
39. Herbert A, Gerry NP, McQueen MB, et al. 2006. A common genetic variant is associated with adult and childhood obesity. *Science* 312:279–83
40. Myers RH. 2006. Considerations for genomewide association studies in Parkinson disease. *Am. J. Hum. Genet.* 78:1081–82
41. Lyon HN, Emilsson V, Hinney A, et al. 2007. The association of a SNP upstream of INSIG2 with body mass index is reproduced in several but not all cohorts. *PLoS Genet.* 3:e61
42. Arking DE, Pfeufer A, Post W, et al. 2006. A common genetic variant in the NOS1 regulator NOS1AP modulates cardiac repolarization. *Nat. Genet.* 38:644–51
43. Dewan A, Liu M, Hartman S, et al. 2006. HTRA1 promoter polymorphism in wet age-related macular degeneration. *Science* 314:989–92
44. Duerr RH, Taylor KD, Brant SR, et al. 2006. A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314:1461–63
45. Hunter DJ, Kraft P. 2007. Drinking from the fire hose—statistical issues in genomewide association studies. *N. Engl. J. Med.* 357:436–39
46. Sladek R, Rocheleau G, Rung J, et al. 2007. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 445:881–85
47. Yeager M, Orr N, Hayes RB, et al. 2007. Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* 39:645–49
48. Rioux JD, Xavier RJ, Taylor KD, et al. 2007. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* 39:596–604
49. Fellay J, Shianna KV, Ge D, et al. 2007. A whole-genome association study of major determinants for host control of HIV-1. *Science* 317:944–47
50. Pearson TA, Manolio TA. 2008. How to interpret a genome-wide association study. *JAMA* 299:1335–44
51. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. *Genet. Med.* 4:45–61
52. Chanock SJ, Manolio T, Boehnke M, et al. 2007. Replicating genotype-phenotype associations. *Nature* 447:655–60
53. Gorlov IP, Gorlova OY, Sunyaev SR, et al. 2008. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* 82:100–12
54. Mailman MD, Feolo M, Jin Y, et al. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* 39:1181–86
55. National Cancer Institute. *Cancer Genetic Markers of Susceptibility (CGEMS) data portal*. **https://caintegrator.nci.nih.gov/cgems/**, accessed 4/29/08
56. European Genotype Archive. **http://www.ebi.ac.uk/ega/page.php?page=home**, accessed 4/29/08
57. Centers for Disease Control and Prevention. *HuGE Navigator*. **http://www.hugenavigator.net/**, accessed 9/12/2008
58. Homer N, Szelinger S, Redman M, et al. 2008. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* 4(8):e1000167
59. Couzin J. 2008. Genetic privacy. Whole-genome data not anonymous, challenging assumptions. *Science* 321(5894):1278
60. Cooper RS. 2003. Gene-environment interactions and the etiology of common complex disease. *Ann. Intern. Med.* 139:437–40
61. Guttmacher AE, Collins FS, Carmona RH. 2004. The family history—more important than ever. *N. Engl. J. Med.* 351:2333–36
62. Moffatt MF, Kabesch M, Liang L, et al. 2007. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* 448:470–73

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

Annual Review of Medicine

Volume 60, 2009

Contents

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.

Indexes

Errata

An online log of corrections to *Annual Review of Medicine* articles may be found at http://med.annualreviews.org/errata.shtml

Annu. Rev. Med. 2009.60:443-456. Downloaded from arjournals.annualreviews.org
by Stanford University - Main Campus - Robert Crown Law Library on 01/02/10. For personal use only.