Chapter 14

Genome-Wide Association Study in Humans

J. Gustav Smith and Christopher Newton-Cheh

Abstract

Genome-wide association studies have opened a new era in the study of the genetic basis of common, multifactorial diseases and traits. Before the introduction of this approach only a handful of common genetic variants showed consistent association for any phenotype. Using genome-wide association, scores of novel and unsuspected loci have been discovered and later replicated for many complex traits. The principle is to genotype a dense set of common genetic variants across the genomes of individuals with phenotypic differences and examine whether genotype is associated with phenotype. Because the last common human ancestor was relatively recent and recombination events are concentrated in focal hotspots, most common variation in the human genome can be surveyed using a few hundred thousand variants acting as proxies for ungenotyped variants. Here, we describe the different steps of genome-wide association studies and use a recent study as example.

Key words: Genome-wide association study, GWAS, whole-genome association study, WGAS, complex genetics, common variation.

1. Introduction

Genome-wide approaches offer a systematic analysis of genes with and without a priori evidence for involvement in the molecular basis of a trait or disease. The benefits of genome-wide studies, as opposed to candidate-gene-based studies, stretch beyond identifying a genetic basis for trait differences to uncovering novel molecular pathways and interactions that underlie the trait, leading to new insights into physiology and pathophysiology. Moreover, the in vivo relevance of a gene to a trait, having been established in humans at the outset, can guide development of model organisms and systems with direct applicability to human disease. Lastly, identified genes may be of special interest to the medical community as a means to predict disease risk at the population level and as novel therapeutic targets.

K. DiPetrillo (ed.), *Cardiovascular Genomics*, Methods in Molecular Biology 573,
DOI 10.1007/978-1-60761-247-6_14,

Genome-wide approaches were first proposed by Botstein and colleagues (1). The principle is to genotype a number of common genetic variants spaced across the genome which can act as proxies for ungenotyped variants due to coinheritance of neighboring variants on a chromosomal segment. Botstein suggested using a panel of restriction fragment length polymorphisms (RFLPs) to examine how familial transmission of allelic variants tracks with transmission of phenotype. When a polymorphism shows significant linkage to a disease, additional markers can be genotyped in that region, termed fine-mapping, and sequencing can be performed to identify the responsible gene and variant. This genome-wide approach examining familial transmission, termed linkage analysis, has been successfully utilized to identify the rare variants underlying more than 2,000 monogenic disorders to date (http://www.ncbi.nlm.nih.gov/Omim/mimstats.html). Monogenic (i.e., Mendelian) disorders, such as familial hypercholesterolemia, are relatively less common in the population and are typically caused by rare variants, often leading to altered amino acid sequence, which have strong (nearly deterministic) effects producing overt disease. Complex traits, by contrast, result from multiple genetic and environmental contributors and are typically much more common, such as hypercholesterolemia. As predicted by Risch and Merikangas, linkage has largely failed to identify common variants in common, multifactorial traits due to inherent methodological limitations (2).

In the last year we have witnessed the success of genome-wide association (GWA) studies, proposed by Risch and Merikangas in the same seminal perspective (2), in identifying dozens of common genetic variants associated with common diseases and traits. These discoveries have provided new insights into the genetic architecture of common traits by providing evidence that common variation explains at least a portion of the variation in risk of common disease, known as the common disease–common variant hypothesis (3).

Genome-wide association studies make use of abundant and easily genotyped molecular markers, single-nucleotide polymorphisms (SNPs), which are single-base substitutions with minor allele frequencies above 0.01. Fundamentally, such association studies are quite simple and test the hypothesis that allele frequency differs between individuals with differences in phenotype. The facts that the out-of-Africa migration approximately 50,000 years ago, from which all modern non-African human populations were founded, was relatively recent on an evolutionary scale (4) and that recombination across the human genome is highly concentrated in hotspots (5) have resulted in a

high degree of correlation between adjacent variants (termed linkage disequilibrium). This linkage disequilibrium allows the majority of common variation to be assayed with 500,000–1,000,000 correctly chosen proxy SNPs (6). Although proposed more than a decade ago, these studies have only recently been made possible by the completion of the human genome reference sequence (7, 8), the deposit of millions of SNPs into public databases (9), developments in high-throughput genotyping technology (10), and the completion of the Human Haplotype Map (HapMap) (11).

Before the era of GWA studies, only a handful of the thousands of genetic polymorphisms examined in candidate genes showed consistent replication (12, 13). This has been attributed in part to poor study design, genotyping error, and population stratification, but most importantly to overly permissive *p*-value thresholds in the original samples (resulting in false-positive results) and inadequately powered sample sizes in replication cohorts (resulting in false-negative results). To address these problems, a working group assembled by the US National Cancer Institute and the National Human Genome Research Institute suggested a set of standards for performing genetic association studies (14). Here, we follow the framework outlined by the working group, highlight steps where there is emerging consensus, and suggest how to proceed where there are no standards. It should also be stressed that since GWA studies were made possible only recently and have gone through rapid developments, there is little consensus for many steps of the process. We cite the central publications that have shaped current GWA methodology for the interested reader, but cite only a few of the large number of GWA studies published. For a collection of all GWA studies published we refer to the homepage of the NHGRI Office of Population Genomics (http://www.genome.gov/26525384).

A GWA study raises many issues in computational management and statistical analysis of the large amount of genotype data generated, which is reflected in the contents of this chapter. We describe the procedures for (1) careful study design before data collection where the critical issue is to clearly define the trait and reduce the impact of potential biases (2); data acquisition, including DNA genotyping, genotype calling, and the quality control of results (3); statistical analysis; and (4) follow-up of results to validate and replicate interesting loci, identify causal genes and variants, and improve power in meta-analysis. To provide an example, we use the Diabetes Genetics Initiative (15), one of the early GWA studies which helped establish standards and identified problems and advantages of the method.

2. Materials

2.1. A Statistics Software Package

R, SAS, SPSS, Stata, StatView.

2.2. Specific Applications for Statistical Genetics (Suggestions, See Note 1)

1. PLINK
   [http://pngu.mgh.harvard.edu/~purcell/plink/]
   Software for management and analysis of GWA datasets (42).
2. snpMatrix
   [http://www.bioconductor.org/]
   Application for management and analysis of GWA datasets implemented in Bioconductor (16).
3. SAS/Genetics (with SAS by SAS Institute, Inc., Cary, NC, USA)
   Statistical genetics analysis package for SAS.
4. R Genetics
   [http://rgenetics.org/]
   Statistical genetics analysis package for Bioconductor.
5. QTDT family-based association test
   [http://www.sph.umich.edu/csg/abecasis/QTDT]
   Application for family-based association analysis (17).
6. FBAT and PBAT
   [http://www.biostat.harvard.edu/~fbat/default.html]
   [http://www.biostat.harvard.edu/~clange/default.htm]
   Applications for family-based association analysis (18, 19).
7. HaploView
   [http://www.broad.mit.edu/mpg/haploview/]
   Software for analysis of correlation patterns in SNP data and visualization of haplotypes and GWA results (20).
8. EIGENSOFT
   [http://genepath.med.harvard.edu/~reich/Software.htm]
   Software for Principal Components Analysis of population substructures (21).
9. STRUCTURE
   [http://pritch.bsd.uchicago.edu/structure.html]
   Software for structured analysis of population substructures (41).
10. MACH
    [http://www.sph.umich.edu/csg/abecasis/MACH/]
    Software for imputation of SNPs not genotyped from genotyped SNPs using LD data (22).
11. IMPUTE (Marchini 2007)
    [https://mathgen.stats.ox.ac.uk/impute/impute.html]
    Software for imputation of SNPs not genotyped from genotyped SNPs using LD data used by the WTCCC (23).

12. WGAViewer (Ge, 2008)
[http://www.genome.duke.edu/centers/pg2/downloads/wgaviewer.php]
Software for visualization and functional annotation of WGA data (24).
13. ANCESTRYMAP
[http://genepath.med.harvard.edu/~reich/Software.htm]
Software for ancestry mapping (25).

2.3. Webpages

1. UCSC Genome Browser
[http://genome.ucsc.edu/]
Genome browser maintained by the University of California Santa Cruz.
2. Ensembl
[http://www.ensembl.org/]
Genome browser maintained by EMBL-EBI and the Sanger Institute.
3. dbSNP and dbGene
[http://www.ncbi.nlm.nih.gov/sites/entrez]
Databases maintained by the National Center for Biotechnology Information containing data on genes and SNPs deposited by researchers.
4. The Human Haplotype Map
[http://www.hapmap.org/]
Genome browser annotated with results from the HapMap project (6, 11).
5. Genetic Power Calculator
[http://pngu.mgh.harvard.edu/~purcell/gpc/]
Online calculator of power in genetic association studies (26).
6. CaTS
[http://www.sph.umich.edu/csg/abecasis/CaTS/]
Online calculator of power in genetic association studies with multistage designs (27).
7. SNP Annotation and Proxy Search (SNAP)
[http://www.broad.mit.edu/mpg/snap/]
Online application for SNP annotation and proxy searching.

3. Methods

3.1. Study Design

3.1.1. Define the Phenotype: Quantitative and Qualitative Traits

As in all epidemiological studies, the first step is careful trait definition and a decision whether to use a qualitative or quantitative trait (*see* **Note 2**). Quantitative traits generally increase power through larger information content, but also may decrease power if imprecise measurements are made, which can be detectable by low inter-reader

or intra-reader reliability. Certain traits, such as diseases, are by definition qualitative but can be measured quantitatively (i.e., when analyzing survival time to an endpoint). Similarly, quantitative traits can be analyzed qualitatively by applying a threshold value; for example, when comparing individuals in the extremes of a distribution to reduce trait heterogeneity.

When using a qualitative trait with cases and controls (i.e., those with and without type 2 diabetes), trait ascertainment should be performed in such a way as to reduce misclassification bias (incorrect case–control assignment) and selection bias (sample not representative of population). Finally, as in all studies of etiology, it may be important to define the phenotype as precisely as possible regarding fundamental mechanism (if well understood) to increase power by reducing heterogeneity, while retaining simplicity to facilitate replication studies and not reducing sample size excessively. The appropriate balance between very selective entry criteria and maximal inclusiveness to increase sample size needs to be considered on a trait-by-trait basis.

Example Several traits were studied in the Diabetes Genetics Initiative (DGI). Here, we focus on one quantitative trait, low-density lipoprotein cholesterol (LDL), and one qualitative trait, the disease type 2 diabetes (T2D), which are both well-studied phenotypes. T2D was defined according to internationally standardized criteria established by WHO. To reduce etiologic heterogeneity, steps were taken to exclude individuals with type 1 diabetes, which is thought to have a different pathophysiology, and individuals with monogenic diabetes (e.g. maturity onset diabetes of the young). Control subjects were defined as having normal glucose tolerance in testing to exclude diabetes or pre-diabetes. Also, they did not have any first-degree relatives with known diabetes. LDL was estimated using the clinical standard of the Friedewald formula, in which measurements of total blood lipid content are performed and non-LDL lipid types (high-density lipoprotein cholesterol, triglycerides) are subtracted. Measurements were performed on standard clinical assays. Lipid lowering therapy, present in a minority of subjects, was an exclusion criterion.

3.1.2. Estimate the Contribution of Genetic Variation to Phenotypic Variation

It is likely that most human traits that exhibit variability have a genetic component. However, it is well known that the degree of heritability differs greatly between traits and heritability reflects the total effect of genetic variation as well as environmental and behavioral factors segregating in families. Hence, it can be useful to assess the amount of trait variability explained by heritability before embarking on a costly genome-wide association study. That is, if the heritability of a trait is low, problems with trait definition, measurement error, or other systematic problems may bias genetic association studies to the null. In such cases, attempts could instead be made to enrich for genetic effects by examining

precisely defined subphenotypes (e.g., measures of insulin resistance or beta cell function as opposed to the clinical definition of type 2 diabetes).

To examine whether evidence of heritability exists, perform a critical literature review and perform heritability analyses in the sample used if family data exist. Studies that examine heritability typically examine familial aggregation, concordance between monozygotic as compared to dizygotic twins (MZ >> DZ concordance suggests a genetic basis), and sibling recurrence risk. Heritability analysis is well-discussed elsewhere (28).

Example Both T2D and LDL are well known to have substantial heritable components. A literature review reveals that T2D heritability estimates vary widely among studies, but most estimates range between 30 and 70%. Sibling recurrence risk ratio is estimated to be relatively high at $\lambda_s \sim 3.0$ and monozygotic twins have higher concordance than dizygotic twins, consistent with a strong genetic basis. A literature review of LDL reveals heritability estimates in most studies around 50% and that monozygotic twins have higher concordance than dizygotic twins.

Estimation of trait heritability was not performed in the DGI population given the absence of adequate family data.

3.1.3. Determine Sample Structure: Case–Control, Cohort, Family-Based, Isolates

The four most common sample structures used are case–control studies, cohort studies, family-based studies, and studies in population isolates. Generally, the choice of sample structure should be based on the prevalence and familial segregation of the phenotype of interest (*see* **Note 2**).

Case–control collections are the most widely used samples since they are comparatively easy and inexpensive to collect, but this design is also the most sensitive to different forms of bias, including selection bias and confounding by ancestry (population stratification). To avoid problems of population stratification, controls should always be selected from the same population as cases. Population stratification can be adjusted for in the data analysis, but steps should be taken to minimize the potential problem up front in the study design. Further matching for sex and other dichotomous covariates with a strong impact on phenotype can also be performed to increase power, but by reducing non-genetic contributions to trait variation. Further gain in power may result from a focus on cases with a family history of a condition, called enrichment sampling, but this remains to be proven. The ratio of controls to cases is a matter of availability of resources. If one has a choice between increasing the ratio of controls to cases versus keeping the ratio at 1:1 but increasing the number of case–control pairs, it is more powerful to increase the number of case–control pairs in 1:1 matching. However, for a trait such as sudden cardiac death, in which cases are not necessarily available to

increase the number of case–control sets, increasing the number of controls to two to three times the number of cases provides some gain in power. Beyond a ratio of 3:1, there is minimal incremental power and it is typically not advisable given the fixed cost of genotyping platforms. If existing and appropriate control samples are available (without additional genotyping cost) there is no downside to increasing the number of controls.

Cohort studies are more difficult and costly to collect since they involve phenotyping a large number of individuals, often with lengthy follow-up to detect incident events. They have the advantage of being more representative of the population and hence less prone to selection bias, but the trait under study must be relatively common in the population for a well-powered GWA study, rendering this study design impossible for rare diseases. While the lack of selection on trait status may reduce potential biases, it may also reduce the precision of a trait measurement compared to a study focused explicitly on a single trait. One advantage is that individuals can be examined and followed for large numbers of traits as exemplified by the multiple GWA studies performed in the Framingham Heart Study (29).

If the phenotype is qualitative, relatively rare, and shows marked segregation in families, a family-based approach is likely to be of great value. However, familial aggregation may be due to shared environmental factors and rare genetic variants. Family-based approaches focus on transmission of alleles from heterozygous parents to affected individuals more often than expected by chance. Methods exist to study only this within-family component, a design with the benefits of not requiring phenotypic information in the parents and resistance to population stratification, but with power reduced by the requirement for informative parent–offspring trios. In addition, methods have been developed to incorporate both the within-family and between-family (as in studies of unrelated individuals) components with a gain in power; the cost is greater susceptibility to population stratification, but current methods to adjust for population structure render this less of an issue. Family-based studies are particularly advantageous in pediatric diseases, with the major disadvantage being the difficulty of ascertaining familial genotypes. In addition, all studies of familial transmission are more sensitive to false-positive association due to genotyping errors than studies of unrelated individuals (30).

Population isolates may have experienced strong founding effects, in which genetic variation carried into an isolate from the source population is reduced and linkage disequilibrium increased. Such samples have proven to be valuable tools in identifying genes for Mendelian traits using linkage analysis, due to the low genetic and environmental heterogeneity. Their value in genome-wide association studies is less clear since the sample sizes are typically small, comparability with other populations is uncertain, and excess relatedness introduces analytical difficulties.

Example In DGI, a case–control approach was used with both population-based and family-based subsamples that were first analyzed separately and then pooled for joint analysis. In the unrelated subsample, cases and controls were matched on gender, age of diabetes onset, collection locale, and body mass index. Up to two cases were matched to up to two controls.

In the family-based subsample one or two discordant sib pairs were included per family. The samples were collected in a number of towns in Finland and Sweden.

3.1.4. Calculate Power and Determine Sample Size

One of the lessons learned from the first wave of GWA studies is the need for large sample sizes. The power to detect an association with sufficient statistical support to distinguish a result from a chance finding depends on sample size, allele frequency, effect size, and coverage of genetic variation on the genotyping array. For common variants, the theoretically expected effect sizes (2) are genotypic relative risks below 2.0, which is consistent with effect sizes seen in most GWA studies (typically 1.1–1.5). Minor allele frequencies examined in most GWA studies range between 0.01 and <0.50, but power is much lower to detect effects at the low end of the frequency spectrum. The genomic coverage reflects the proportion of all genetic variants for which the genotyped variants alone or in aggregate (as haplotypes) act as proxies at a certain level of linkage disequilibrium measured as the coefficient of determination, r^2, the proportion of variation of SNP explained by another SNP.

Since the effect estimates seen for SNPs in GWA studies are typically modest, it follows that sample sizes have to be very large in order to achieve p-values below the significance thresholds currently recommended (5×10^{-8} to 1×10^{-7}; *see* **Section 3.3.2**). As a guideline, for SNPs with minor allele frequency >0.2 sample sizes of 1,000 cases and 1,000 controls are required for odds ratios of 1.5 with 80% power. Estimates of the sample size needed for certain levels of power can be obtained using the online applications Genetic Power Calculator or CaTS (*see* **Section 2**).

Genome-wide association analysis can be performed in two stages to reduce genotyping costs while retaining power (27). A large number of markers are first genotyped in a random subset of a sample and the strongest associations are then genotyped in the full sample, reducing costs compared to genotyping all markers in the entire sample. However, arrays with fixed marker sets are becoming less costly and may soon be more cost-effective than genotyping large numbers of individual markers. Power in one- and two-stage designs can be estimated using the software CaTS (*see* **Section 2**), but in general favor somewhat larger second-stage samples.

Example Using CaTS, Zondervan and Cardon estimated that a sample size of about 6,500 cases and 6,500 controls was needed to detect association with a SNP conferring a relative risk of 1.2 and an allele frequency of 0.2, assuming 300,000 independent tests at 80% power at a significance threshold of 1×10^{-7} (31).

3.2. Data Acquisition and Quality Control

3.2.1. Genotyping and Genotype Calling

Several high-throughput genotyping platforms assaying marker sets of different density and selection method are currently available, mainly from three companies: Affymetrix, Illumina, and Perlegen. Marker densities and selection methods correspond to different coverage of genome-wide variation, which is a determinant of the power to detect a true association if sample size is adequate (32). Marker density currently ranges from 100,000 to 1,000,000 variants per array selected using one of four methods: (a) randomly; (b) through a tagging approach to maximize coverage based on linkage disequilibrium patterns; (c) a combination of the first two; or (d) focusing only on SNPs known or likely to be functional. The SNPs on most arrays are a mix of variants that are polymorphic in populations of differing ancestry such that a number of SNPs will always be uninformative because they are monomorphic or rare in the population genotyped. Generally, the choice of array should focus on maximizing power taking into account parameters of sample size, ancestral origin, and marker selection methods. With adequate sample size, increased marker density increases power up to a threshold of about 500,000 markers. One million SNPs are likely to be of only modest value in populations other than those of African ancestry, in which linkage disequilibrium is slightly lower and genetic diversity slightly higher. Power calculations for different arrays have been published (33–35) and currently available arrays have recently been reviewed (*see* **Notes 3 and 4**) (10).

Sample collection and handling should be as uniform as possible to minimize bias from batch effects from different sources of DNA, extraction protocols, genotyping procedures, and plate effects. To avoid plate effects (e.g., poor genotyping specific to an individual plate), assignment of individuals to genotyping plates should be randomized blinded to case–control status, while preserving case–control clusters within plates to avoid differential effects within case–control clusters across plates.

A number of different algorithms for allelic discrimination from the intensity data generated in the genotyping procedure, termed genotype calling, have been proposed. The basic principle is clustering of results with similar signal intensities into discrete genotypes. Most genotyping assays come with a recommended calling algorithm. For example, Affymetrix currently recommends the Robust Linear Modeling using Mahalanobis distance (RLMM) procedure described by Rabbee and Speed (36) and developed further by

inclusion of a Bayesian estimation component, called BRLMM. Other well-known calling algorithms include CHIAMO, developed by the WTCCC, and Birdseed, developed at the Broad Institute.

Example In DGI, genotyping was performed on the Affymetrix Human Mapping 500 K GeneChip, which contains oligonucleotides to assay 500,568 SNPs. A total amount of 1 μg of genomic DNA from cases and controls were interspersed equally across 96-well plates. Genotyping was performed according to recommendations from the manufacturer. Using BRLMM, a call rate of 99.2% was achieved for 386,731 SNPs passing quality control.

3.2.2. Genotype Quality Control

Genotyping errors are a potential cause of spurious associations and must be carefully sought and addressed, both on the DNA and SNP levels.

On the sample side, potential errors to assess include inadvertent sample duplication, contamination, and low DNA quality. An efficient way to address duplications and sample contaminations is to compare genome-wide SNP data between each pair of individuals in the sample using the population genetic tools of identity-by-descent (IBD) probabilities and coefficient of inbreeding estimates (F). IBD probabilities are calculated from pairwise identity-by-state (IBS) distances, which can also be used to assess population substructure and cryptic relatedness (discussed later in **Section 3.2.4**). IBD probabilities for each pair of individuals can be assessed using the Pihat method, which estimates the proportion of the genomic variation shared IBD. Identical twins, and duplicates, are 100% identical by descent (Pihat=1.0), first-degree relatives are 50% IBD (Pihat=0.5), second-degree relatives are 25% IBD (Pihat=0.25), and third-degree relatives are 12.5% equal IBD (Pihat=0.125). F can be calculated from homozygous genotypes using PLINK. Whether duplications or twins, most studies will randomly include only one individual from pairs of Pihat 1.0. Contamination can be detected by a pattern of low-level relatedness to many people, a strong negative F, and an elevated genome-wide heterozygosity estimate. Individuals with a high rate of missing genotypes should be excluded, as this is an indication of low DNA quality. Commonly used approaches involve exclusion of all individuals with <2–5% missing genotypes, depending on the platform and sample quality. As an additional sample quality control check, most arrays offer a sex confirmation assay that should be compared with self-reported gender.

On the SNP level, errors in genotyping need to be assessed. A good indicator of genotype quality is the SNP call rate, so SNPs with a high proportion of missing genotypes should be excluded. Commonly used thresholds range from 2 to 5% missing genotypes. SNPs with a low population frequency (minor allele frequency, MAF) may also be excluded, as power to detect association with

these is low and genotype calling is more prone to errors, especially for case–control analysis. A commonly used threshold is to exclude SNPs with minor allele frequency below 1–5% depending on the sample size (larger sample sizes have better power at lower frequencies). Another quality control test for genotyping errors is deviation from Hardy–Weinberg equilibrium (HWE). The benefits of excluding SNPs that fail an HWE test should be carefully weighed as disequilibrium typically is underpowered for genotyping error detection, but may be caused by population stratification, excess relatedness, positive or negative selection, or true association signals. If HWE is used, it should only be calculated in controls if possible by applying stringent thresholds, in the range of 10^{-6}, as true association signals in merged case–control data can generate only modest deviations from HWE (37, 38). HWE can be calculated using Pearson's Chi-square test or Fisher's exact test, as implemented in most statistical analysis applications. A quantile–quantile plot can be useful to determine whether general HWE disequilibrium appears likely to be caused by population stratification, excess relatedness, selection, or association signals in the same manner as described in **Section 3.2.4** for association test statistics. An additional genotyping quality control sometimes used is deliberate duplication of a number of individuals for comparison of concordance rates.

In family-based studies, an additional quality-control filter is to exclude SNPs which fail to show transmission according to Mendelian inheritance laws. The threshold for Mendelian error filtering of DNA samples or SNPs is based on the number of possible errors given the number of related individuals and number of SNPs examined.

After genotype quality control, we recommend examining whether genotyping failure rates differ between cases and controls using a Pearson's Chi-square test with one degree of freedom to exclude false associations due to systematic differences in quality of genotyping or DNA that can lead to spurious association (39). As a final quality control of SNP genotype calling, visual inspection of cluster plots should be performed after analysis for associations that are considered significant (14).

Example In the DGI analysis, individuals with genotyping call rates below 0.95 were excluded. IBD probability estimates from PLINK were used to verify reported familial relationships. Individuals in the whole sample with cryptic (unrecognized) first-degree relatedness were excluded. Individuals whose sex based on genotypes was discrepant from that self-reported were also excluded. After these exclusions, genotypes for 2,931 individuals remained.

Exclusion criteria for SNPs in DGI included: mapping to multiple locations in the genome (3,605 SNPs), call rate < 0.95 (34,532 SNPs) or < 0.90 in either the family-based or the

population-based sample (229 SNPs), MAF < 0.01 (66,787 SNPs) or MAF < 0.01 in either subsample (2,909 SNPs), and Hardy–Weinberg disequilibrium in controls with $p < 1 \times 10^{-6}$ (5,775 SNPs). After quality-control filters were applied, genotypes for 386,731 SNPs were obtained. Testing for differences in missing data between cases and controls was not significant.

3.2.3. Phenotype Quality Control

As with all datasets, quality control must be performed on the phenotype dataset. All individuals should match predefined criteria for inclusion and exclusion. All categorical variables should be examined for improperly coded results, and continuous data should be checked for outliers with consideration to exclude individuals with extreme values. Continuous variables should also be assessed for relative normality in distribution. Helpful tools to this end are visual inspection of a histogram and measures of skewness and kurtosis. If a right-skewed distribution is observed, a logarithm-transformation is often performed to normalize the dataset.

Example The absence of individuals with exclusion criteria and presence of individuals with inclusion criteria was confirmed. No aberrant categories in T2D were detected. LDL values exceeding the 99.5 percentile (>7.423 mmol/L) were excluded as outliers (n=15 individuals) and 197 individuals receiving lipid-lowering therapy were excluded. The LDL distribution was considered near-normal upon inspection of the histogram, examination of skewness and kurtosis, and comparison of mean and median.

3.2.4. Handling Population Stratification and Relatedness

Population stratification, also called confounding by ancestry, arises when samples are composed of subgroups with differences in allele frequencies, due to either different ancestry or excess relatedness, leading to phenotypic differences. This can confound a study, resulting in both false-positive and false-negative findings (39, 40). For example, in Campbell et al. (40), the coincidence of gradients of high to low frequency in the lactase persistence allele and of high to low height from northern to southern Europe was shown to result in spurious association of height with the lactase persistence allele. Hence, it is recommended to test for population structure in samples using several tools (*see* **Note 5**). A high degree of shared ancestry and relatedness in the study sample makes a family-based study design more tractable (discussed in **Section 3.3.2**).

Of the several tests developed to examine population stratification using genome-wide SNP data, we recommend first estimating the extent of association test statistic inflation using a quantile–quantile (QQ) plot for both HWE and association tests. In these plots, observed p-values for each SNP are ranked in order from smallest to largest and plotted against expected values from a theoretical distribution, such as the χ^2 distribution. Population stratification can be seen as a global excess of higher observed

p-values than expected throughout the span of the plot. This stems from a simplifying hypothesis that only a modest number of variants are associated with a given trait and association of thousands of SNPs indicates population stratification. In addition, a quantitative measure of this deviation can be derived from the genomic inflation factor, λ_{GC}, which is easily calculated as the ratio of the median association test statistic over the theoretical median test statistic of a null distribution (for χ^2 this is 0.675^2) for one degree of freedom tests (41). If using a genotype test – in which each genotype class *AA*, *Aa*, *aa* is tested for a difference in phenotype – the procedure accommodates the two degree of freedom test (42). Under a completely null distribution, observed *p*-values correspond exactly to the expected λ_{GC} of 1. A λ_{GC} that exceeds 1 represents global inflation of the test statistic and can arise from biased genotyping failures between cases and controls or, more commonly, population stratification not accounted for in the genotype–phenotype procedure. Such population stratification can be corrected for by dividing the χ^2 statistic of each SNP by the λ_{GC} value, which makes less extreme all observed *p*-values and ensures that the mean is one, a procedure termed genomic control (41). Genomic control may increase the risk of a false-negative association (type 2 error) somewhat because of its global approach to adjusting *p*-values, but can increase power overall by enabling the use of different, stratified samples. An alternate, less commonly used approach is to calculate λ_{GC} from a number of "null" SNPs not expected to show an association with the trait under study.

One can also detect differences in ancestry among individuals in a study and adjust the analysis using any of the following approaches: a principal components analysis approach called the EIGENSTRAT method implemented in EIGENSOFT (21); examination of the coefficient of relationship (*F*); a structured approach (43); and an approach based on identity-by-state distance and multidimensional scaling implemented in PLINK (44). If a subset of individuals with ancestry that is distinguishable from the overall is identified, they can be excluded or, if clustering into several subgroups, can be analyzed together in a stratified analysis such as the Cochran–Mantel–Haenszel test (discussed in **Sections 3.3.2** and **3.4.3**).

Example To account for the excess relatedness in the family-based subsample of the Diabetes Genetics Initiative, association testing for T2D was performed separately in the family-based sample and the unrelated sample after which the results were pooled in a meta-analysis as described in **Sections 3.3.2** and **3.4.3**.

To examine population stratification and excess relatedness, *p*-value distributions were first examined in QQ-plots and genomic inflation factors were calculated. The genomic inflation factors were λ_{GC}=1.05 for T2D and λ_{GC}=1.07 for LDL, corresponding

to very mild overdispersions of test statistics as compared to the null distribution. Next, population stratification was assessed using the EIGENSTRAT method implemented in EIGENSOFT. After clustering outlier samples identified by EIGENSTRAT and meta-analysis using a Cochran–Mantel–Haenszel stratified test, *p*-values were very similar to *p*-values attained by genomic control as evidenced by r^2=0.95. Hence, genomic control was used to adjust for population stratification in both traits because of its simplicity.

3.2.5. Imputation of Ungenotyped SNPs

Genotypes of untested SNPs can be imputed from genotyped SNPs using LD patterns from the International Human Haplotype Map (HapMap) to expand coverage over the 3 million HapMap SNPs and allow meta-analyses across platforms using different marker sets. To date, the value of imputation in increasing power has not been clearly demonstrated, but imputation has successfully facilitated several meta-analyses using fixed genotyping arrays with different coverage. Several applications for imputation, using different methodologies, have been developed, including MACH, IMPUTE, and a method implemented in PLINK (*see* **Section 2**).

Quality-control metrics for imputation provided by the imputation applications are generally a function of the ratio of the observed to the expected variance on the imputed allele count. The ability to impute a genotype for a given SNP is a function of the accuracy of the directly genotyped SNPs in the region and their correlation to the untyped SNP. Note that the increased number of tests with imputation does not demand additional correction for multiple testing if using a significance threshold based on the total number of independent tests in the human genome (*see* **Section 3.3.2**).

Example Untyped HapMap phase II SNPs were imputed in the DGI population to increase power and allow pooling of data from different marker sets in a meta-analysis with other genome-wide datasets of T2D (45). A total of 2,202,892 SNPs were imputed using MACH and passed quality control across all three component studies. DGI and WTCCC both used the Affymetrix Human Mapping 500 K (22), and the FUSION study used the Illumina HumanHap300 BeadChip (46). SNPs with an estimated MAF<0.01 in the pooled or any individual sample were excluded. The imputed SNPs showing the strongest association were directly genotyped in DGI using a different method for validation and showed good agreement with imputed genotypes (r^2=0.84).

3.3. Statistical Analysis

3.3.1. Phenotype Modeling

The goal of phenotype modeling is to increase power through reduction of non-genetic contributions to trait variation by adjusting the trait of interest for variation attributable to known covariates. Generally, it is difficult to establish whether covariates are

etiologically independent from the genetic pathways sought, but most continuous traits vary with age and sex and are often adjusted for these variables. Age, in particular, was recently found to modify the effect of association findings (47). A thorough literature review should be conducted to identify potential covariates for adjustment. In general, it is desirable to include covariates that are reproducibly measured, available in other potential replication cohorts, and are unlikely to be in the causal pathway between genotype and disease outcome.

For continuous traits, adjusted models can be created through regression modeling with univariate and stepwise regression. The proportion of variability explained by a model can be estimated as the coefficient of determination, r^2. Residuals (the observed trait value minus that predicted from a regression model) can be used as the trait for genotype–phenotype analysis in the GWA study.

Example LDL cholesterol was adjusted for covariates with previous evidence, high reproducibility, and significant association using multivariable linear regression modeling. Included variables were age, age^2, sex, enrollment center, and diabetes status. Residuals were standardized to a mean of 0 and a standard deviation of 1.

T2D was not adjusted for any covariates since individuals with other recognized etiologies were excluded in the design stage and cases and controls were matched on age, sex, and body mass index.

3.3.2. Association and Statistical Inference

The association between phenotype and genotype is examined by comparing allelic patterns between individuals with different phenotypes. All inheritance models can be examined (additive, dominant, recessive) in separate allelic tests with one degree of freedom, which improves statistical power to detect effects that follow one model more than another, but increases the number of hypotheses tested. Hence, most studies of qualitative traits with unknown allelic inheritance patterns have used either a general genotype test or an additive (allelic trend) test. The general genotype test compares distribution among frequencies of the three genotypes (minor homozygous, major homozygous, heterozygous) between cases and controls and most often uses a Pearson's Chi-square test with two degrees of freedom or Fisher's exact test. The allelic trend test assumes an additive genetic model and tests whether the slope of a fitted regression line, with the three genotypes as independent variables, differs from zero and can be adjusted for covariates. The trend test most often used is the Cochran–Armitage test, whereas Pearson's Chi-square test with one degree of freedom for minor allele frequency is not recommended (48).

For quantitative traits, statistical inference is performed using either linear regression, which assumes an additive model, or ANOVA, which is based on a general model. Both tests require

the trait to be approximately normal distributed, which may be achieved by transformation (e.g., $\log_{10}$) if the trait is positively skewed and by adjustment for relevant covariates.

A number of tests for hypothesis testing in family-based samples have been developed. The transmission disequilibrium test (TDT) has one of the simplest designs and examines transmission of qualitative traits within trios (parents and one offspring), with the advantage that only offspring need to be phenotyped. FBAT (family-based association test) implements the TDT for both quantitative and qualitative traits (18). These methods offer the advantage of being resistant to population stratification, but suffer from reduced power because only within-family components are examined and only some parental genotype patterns are informative. Family-based methods with increased power test both within-family and between-family variations for association with genotype, but with a consequence of some sensitivity to population stratification. Such methods are used in DFAM (implemented in PLINK) for qualitative traits, while QTDT (17), QFAM (implemented in PLINK), and GEE (49) model quantitative traits, and PBAT handles both quantitative and qualitative traits. Note that large pedigrees must be broken down to computationally tractable units for these tests. These or related methods have been utilized in recent GWA studies (29, 50–52). Most of the tests described above have been implemented in the statistical analysis applications named in **Section 2**.

The mass significance problem caused by the multiple tests of a GWA study is currently the subject of some debate (*see* **Note 6**). Although Bayesian approaches incorporating information on power and expected number of true positives received much interest early on (53), the frequentist approach of adjusting for a number of independent tests, as originally proposed by Risch and Merikangas (2), is most often used. The Wellcome Trust Case Control Consortium settled on a genome-wide significance threshold of 5×10^{-7}, based on power calculations showing that their study of 2,000 cases and 3,000 controls reached 80% power to detect an association at this significance level for SNPs with MAF>0.05 and relative risk of 1.5, with a drop in power to 43% for relative risks of 1.3 (23). Others have suggested a significance level of 5×10^{-8} in populations of European ancestry based on a overall genome-wide significance threshold of 0.05 adjusted for an estimated 1 million independent SNPs in the genome (11, 54, 55), by the Bonferroni method, paralleling the linkage thresholds proposed by Lander and Kruglyak for genome-wide linkage analysis (56). Independent SNPs for Asian populations sampled in HapMap represent a similar number of independent tests, whereas African-derived samples represent more tests, based on the extent of variation and LD (6, 11).

As discussed in **Section 3.1.4**, the typically small effect sizes seen for common variants necessitate large sample sizes to reach significance thresholds such as 5×10^{-8}. To reduce genotyping costs, a two-stage design is often utilized where genome-wide marker sets are analyzed in a smaller number of individuals in stage 1, SNPs below a less stringent threshold are genotyped in additional individuals in stage 2, and then SNPs are analyzed jointly in all individuals (27).

The effect size of a SNP is calculated as the odds ratio or relative risk per genotype for qualitative traits and as the beta coefficient from a regression model, which can be interpreted as the effect of each additional allele, for quantitative traits.

Example For T2D the population-based sample was analyzed with a Cochran–Mantel–Haenszel stratified test, as implemented in PLINK, based on the subsample matching criteria (BMI, gender, geographic region) as described in **Section 3.1.3**. The family-based sample was analyzed using the DFAM procedure in PLINK conditioning on sibship as strata. *p*-values from the two subsamples were then converted to *Z*-scores and combined with a weighted *Z*-meta-analysis. A significance threshold of 5×10^{-8} was used.

For LDL, residuals from the phenotype modeling described in **Section 3.3.1** were examined for association in PLINK with genotypes using linear regression modeling with genotypes as the independent variable.

The most significant SNP from the T2D analysis was located in *TCF7L2*, which has previously been associated with T2D, and reached a *p*-value of 5.4×10^{-6}. The most significant SNP for LDL was located in the *APOE* gene cluster that is known to influence LDL metabolism ($p=3.4 \times 10^{-13}$ after performing genomic control). The second and third strongest SNPs for LDL reached *p*-values of 2.3×10^{-8} and 8.9×10^{-8}, respectively. Thus, for these traits two SNPs at two loci reached genome-wide statistical significance in the primary analysis of DGI.

3.4. Interpretation and Follow-Up of Results

3.4.1. Validation, Replication, and Extension to Other Populations

Because of the difficulties of specifying a significance threshold in genome-wide association studies and the multitude of potential methodological errors, ultimate proof of association must come from replication of significantly associated SNPs in independent samples (14). Even for SNPs near genes that constitute strong biological candidates, from previous experimental studies or similar Mendelian traits, replication is still necessary for convincing association.

Technical validation refers to reanalysis of associated SNPs on a different genotyping platform and provides evidence that an observed association signal is not due to systematic genotyping errors. To show this, concordance of genotypes between the assays

is calculated. Technical validation may be considered before replication studies are undertaken, but multiple correlated SNPs at a locus with comparable association argue against genotyping artifact as the source of apparent association. A replication study of a putative association tests the same direction of effect for the same polymorphism in the same phenotype and in a sample from the same ancestral origin and similar ascertainment, often termed exact replication. Studies examining the same locus in populations of different ancestral origin are necessary to demonstrate the relevance of a specific allele to individuals of other ancestries. Care should be taken to examine the linkage disequilibrium patterns before specifying SNPs for follow-up in samples of other ancestries. For example, LD breaks down over shorter distances in individuals of African ancestry relative to those of European ancestry based on the HapMap samples. Following up findings from European-derived samples in African-derived samples may require genotyping of additional SNPs that might be highly correlated on European chromosomes but poorly correlated on African chromosomes. This can be determined using any tagging program, such as the Tagger function in HaploView (*see* **Section 2**).

The power of the replication study to confirm the findings in light of the allele frequency and the effect size observed in the original study should be considered. True positive results that just achieve significance (by whatever appropriate threshold) tend to overestimate the true effect size due to random chance. This will hinder replication of a finding in a second sample if the sample is inadequately powered to find a weaker effect. Power and sample size for replication samples can be calculated in the software packages Genetic Power Calculator or CaTS (*see* **Section 2**). Failure to confirm a finding in an adequately powered second sample could also suggest a false-positive result in the original sample. Iterative meta-analysis of results, as replication samples are genotyped, can help clarify whether a signal is getting stronger or weaker with analysis in independent samples. As the meta-analyzed samples get larger, true-positive results will get more significant, while false-positive results will get less significant.

Example One hundred and fourteen SNPs from the extreme tail of *p*-values for T2D in DGI were genotyped with an independent technology for validation. Concordance rates were 99.5%, indicating that genotyping artifacts did not explain the lowest *p*-values observed.

For T2D, the 59 SNPs with the strongest associations were selected for replication and were genotyped in 10,850 individuals from three case–control samples of European ancestry from Sweden, the United States, and Poland. T2D definitions were the same as those used in the original GWA studies. Combined

analysis of all four samples was performed using the Mantel–Haenszel odds ratio meta-analysis. This analysis identified three genome-wide significant SNPs at three loci.

For LDL, the 196 SNPs with the strongest associations were selected for replication in 18,554 individuals of European ancestry from Sweden, Finland, and Norway. LDL was measured using standard methods and the same exclusions and adjustments for covariates as in the GWA study were applied. Association was examined using linear regression modeling as in the GWA study and results were summarized using inverse-variance-weighted meta-analysis with fixed effects. Two loci were replicated. The findings were also examined in a population of Asian ancestry by genotyping in 4,259 individuals from Singapore. Neither SNP was significant in this population, but it was clearly underpowered to detect association.

3.4.2. Identifying the Causal Variant

Since only a fraction of common genetic variation is genotyped and SNPs act as proxies for untyped SNPs and other genetic variants, it is unlikely that a significantly associated SNP is directly causal. Consequently, GWA studies cannot provide unambiguous identification of causal genes, even if an associated SNP is situated close to or in a gene. Indeed, several GWA studies have found strong association signals from genomic regions containing multiple genes of strong biological candidacy, while others have found regions of association without any known genes, possibly reflecting remote regulatory elements, long-range LD, or incomplete gene annotation.

The first step toward identifying the causal variant is to examine whether the SNP showing the strongest association at a locus or any SNP to which it is highly correlated may be a functional SNP (e.g., missense or nonsense). Information on SNP function relative to nearby genes is available in public databases, such as dbSNP, and can be efficiently retrieved using bioinformatic tools such as the WGAViewer (24), SNAP, or a function implemented in PLINK. Typically multiple SNPs at a locus show association and examining the correlation among associated SNPs in HapMap can show whether correlation between SNPs explains the p-values seen across the locus. Some loci have been found to contain several causal variants.

Second, fine-mapping of additional SNPs known to be correlated in HapMap above a specified threshold with the most significant SNP at a locus can be performed both to narrow down the locus for resequencing and to examine correlated functional SNPs. Correlation among SNPs is measured by the coefficient of determination, r^2. Another option is to perform fine-mapping in silico without further genotyping by imputing untyped variants (discussed in **Section 3.2.5**).

Ultimately, resequencing of the associated locus is necessary to identify the set of all potential variants which could explain a SNP's association. Molecular biologic approaches are then required to determine which among them is most likely to be causal (*see* **Note 7**).

Example None of the SNPs for LDL or T2D with significant association were likely to be functional based on encoding an amino acid change when searching GeneCruiser using SNAP (*see* **Section 2**) and are presumed to be in LD with the causal variants. Fine-mapping and resequencing of significantly associated loci are in progress.

3.4.3. Meta-analysis

Meta-analysis is the statistical procedure of jointly analyzing multiple studies to increase power. For GWA studies, this approach has identified additional common variants of smaller effect for several traits, which may still provide new insights into pathophysiology and physiology although the individual predictive values are small.

Before meta-analysis, all studies included should be examined for possible sources of heterogeneity, both in design and results of included studies. Most studies include only individuals of the same continental ancestry (*see* **Note 5**). Summary association statistics from both population-based and family-based subsamples can be pooled (57). Heterogeneity of results between studies can be examined by plots (Forest plot, Galbraith plot, or funnel plot) or with formal tests such as the Cochran's Q and the I^2 statistics (58).

Phenotypes and genotypes should be quality controlled prior to analysis as described in **Sections 3.3.2–3.2.4** before pooling. Meta-analyses are then performed either under the assumption that variation between studies is due to random variation in which large studies are expected to give identical results (termed fixed-effects models) or under the assumption that variation between studies is due to different underlying effects for each study (termed random-effects). When statistical heterogeneity is detected, random-effects models should be used. Under either model, meta-analysis is performed by combining the data after weighting the results from each study. Combined analysis can be performed with effect estimates (odds ratios or beta estimates) weighted by the inverse variance, with p-values weighted using Fisher's-combined method, or by converting p-values to Z-scores based on the p-value and effect direction. For qualitative traits, meta-analysis can also be performed by pooling genotype counts using the Cochran–Mantel–Haenszel stratified method. Pooling of results at the individual level is problematic in primary screens as they increase the probability of false-positive association due to population stratification. However, examination of epistatic or gene–environment interactions can be facilitated by pooling

individual-level data and by performing association analysis across all individuals, but this is conditional on achieving a significant result upon meta-analysis of summary results from separate analyses.

Example Meta-analysis for T2D was performed with two other studies from the United Kingdom (n=1,924 cases and 2,938 controls) and Finland (n=1,161 cases and 1,174 controls), resulting in a pooled sample size of 10,128 individuals (45). Using data from the HapMap CEU sample of European ancestry, 2,202,892 SNPs were imputed using the MACH software package, enabling comparison across samples with greater coverage. Each study was corrected for population stratification, cryptic relatedness, and technical artifacts before meta-analysis. Heterogeneity among studies was examined for SNP odds ratios using Cochran's Q and the I^2 statistics. Meta-analysis was then performed using Z-statistics derived from p-values and weighted so that squared weights sum to 1; each sample-specific weight was proportional to the square root of the effective number of individuals in the sample. p-values for genotyped SNPs were used where available and results for imputed SNP were used for the remainder. Weighted Z-statistics were summed across studies and converted to p-values. Odds ratios and confidence intervals for the pooled sample were obtained from a fixed-effects inverse variance model. Genomic control was performed on the combined results, although the meta-analysis had a λ_{GC} of only 1.04. Population stratification was also examined using principal components analysis, as described in **Section 3.2.4**.

Imputed variants showing association were directly genotyped for validation before replication of all SNPs in two-replication stages with a total sample size of 53,975 individuals. Based on these analyses, six novel loci were identified that met a genome-wide significance of $p<5 \times 10^{-8}$, most of them with very small effect sizes (odds ratios between 1.1 and 1.2), although effect sizes for the actual causal variants are likely to be larger.

Meta-analysis for LDL was performed with two studies from Finland (n=1,874) and Sardinia (n=4,184), resulting in a pooled sample size of 8,816 individuals (59, 60). Imputation was performed as described for T2D. Analyses were performed using linear regression modeling in each sample and then analyzed jointly using inverse variance-weighted meta-analysis with fixed effects. Heterogeneity was examined in the same way as for T2D; 30 SNPs, representing novel loci with $p<10^{-4}$ from the meta-analysis, were selected for replication in 18,554 individuals. This pooled analysis identified five additional loci associated with LDL.

4. Notes

1. As can be seen in **Section 2**, several applications of high quality for management and analysis of data have been developed by different groups. For newcomers to the field, we recommend the software PLINK, which can be downloaded freely, is regularly updated by leaders in the field, has a logical structure, and has thorough documentation provided on its webpage. PLINK output files can also be visualized in HaploView.
2. Ideally, the study design should be carefully planned before sample collection and genotyping as discussed in **Section 1**. In reality, samples collected previously for other purposes have been used in most genome-wide association studies. We find that the issues discussed in the sections on quality control of genotype and phenotype data as well as population stratification may increase power in these samples. Heritability estimates and coefficients of determination for trait models can be useful indicators of data quality. Most often, the sample available is also the major determinant of whether a case–control, cohort, or family-based design is to be implemented.
3. GWA studies remain an expensive undertaking, but prices are dropping, especially as whole-genome resequencing studies draw near. To reduce genotyping costs, we recommend considering three strategies. First, instead of selecting an expensive array with high coverage, invest in increasing sample size using more arrays with moderate coverage. Coverage can then be increased using imputation procedures (61). Calculations of cost efficiency can be performed using online tools (http://www.well.ox.ac.uk/~carl/gwa/cost-efficiency). Second, the two-stage approach discussed in **Section 3.1.4** can be used. Third, DNA pooling can be performed when genotyping to reduce costs. However, this approach suffers from reduced power and genotype quality-control difficulties (62).
4. Genome-wide SNP data can also be used for other analyses, bringing additional value to the high costs of these analyses. One such analysis is homozygosity mapping, in which low-frequency recessive variants can be examined for trait association. The method involves identification of segments of homozygosity in genome-wide SNP data using hidden Markov models or a method implemented in PLINK (63). These segments can then be examined for trait association. Similarly, haplotypes can be inferred from SNP data and haplotype association tests performed.

Another analysis possible for traits caused by genetic differences between ethnicities is admixture mapping, which can be used to identify the genetic determinants for these traits in admixed populations (25, 64).

Finally, it has been suggested that pathway-based association analyses, similar to the GSEA method used in gene expression array analyses, may be able to identify additional variants (24).

5. We recommend that individual genome-wide association studies be performed in populations of similar ancestry. It is known, through the HapMap project and studies of population genetics, that allele frequencies, and to a lesser extent LD patterns, differ between ancestries. Also, founder populations are known to differ from other populations. Even after conditioning on a different continental ancestry, assessment of population structure using the approaches mentioned in **Section 3.2.4** is highly recommended to avoid false-positive signals (39, 40). The appropriateness of meta-analyzing results across samples of differing ancestry remains unclear, but it may be reasonable to do so if similar correlation patterns exist.
6. GWA studies have been finding reproducible loci associated with complex traits and have largely avoided the flood of non-replicable findings that were seen in candidate gene-based association studies (12) and when genome-wide linkage analysis was first applied to complex diseases (65). This is likely due to the large sample sizes examined, careful study design and quality control, stringent thresholds for statistical significance, and use of independent samples for validation and replication (66).

   There is no doubt that large samples are necessary, that study design and quality control is paramount to reduce bias, and that replication and validation are key to identify significant associations, but the problem of correcting for multiple hypotheses remains a subject of debate. Several approaches for multiple hypothesis correction other than the frequentist Bonferroni and Sidak corrections have been proposed, including the False Discovery Rate (67), which controls the false-positive rate, and Bayesian approaches such as the False Positive Report Probability (53), which determines the probability of an association being falsely positive based on *p*-value, power, and estimates of prior probability. Other approaches include permutation of phenotypes or genotypes to empirically determine the nominal *p*-value that corresponds to a study-wide *p*-value of 0.05. Until these approaches have become established and validated we recommend using the stringent threshold of 5×10^{-8} based on a an overall genome-wide significance threshold of 0.05 corrected by the Bonferroni method for

estimates from HapMap of 1 million independent tests (11, 54, 55), in populations of European and Asian ancestries. The number of independent tests in African populations is likely to be higher.

7. For the hundreds of loci that have been discovered to date, the road from SNP identification to biology is long. As discussed above, identifying the causal variant requires resequencing of the locus in most cases. Furthermore, definitive identification of the gene affected by the causal variant requires functional studies, as genetic variants can be located in regulatory sites and affect distantly located genes as shown in the ENCODE project (68). Functional analyses may prove difficult because variants can affect gene products in multiple ways, including changes in protein sequence or changes in expression due to different gene copy number, regulation of gene transcription, and regulation of RNA post-processing (capping, splicing, polyadenylation). To date, databases of association of SNPs with transcriptome-wide expression patterns (69, 70) and alternative splicing (71) in specific cell lines are available.

   Databases of copy number polymorphisms (CNPs), another form of genetic variation in LD with SNPs that has recently been shown to have a large impact on interindividual genetic variation, are also available and the most recent microarrays include assays for large numbers of CNPs (72). The relative importance of different genetic variants such as SNPs and CNPs, or for that matter common and rare variations, is currently unknown, but will be elucidated once resequencing of genomes becomes more common. This will enable characterization of the full spectrum of genetic variation that contributes to interindividual differences for any given trait.

   Genetic variants may also act in concert with each other, termed epistasis, or with environmental factors to cause trait changes. Studies of such interactions require even larger sample sizes than GWA studies and are likely to follow from the current collaborative projects of pooling datasets.

8. For more information on general methodology see excellent recently published reviews (48, 62, 73–75). For specific points on study design see a recently published protocol (31) and for more information on the history of the population genetic theory leading to GWA study approaches see a recent discussion (76). For examples of genome-wide association scans using different methodologies we refer to the collection of all published studies maintained by the NHGRI office of population genomics at their homepage (www.genome.gov).

References

1. Botstein, D, White, RL, Skolnick, M, et al. (1980) Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet* **32:** 314–331.
2. Risch, N, Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* **273:** 1516–1517.
3. Lander, ES. (1996) The new genomics: global views of biology. *Science* **274**: 536–539.
4. Reich, DE, Cargill, M, Bolk, S, et al. (2001) Linkage disequilibrium in the human genome. *Nature* **411:** 199–204.
5. McVean, GA, Myers, SR, Hunt, S, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* **304:** 581–584.
6. Frazer, KA, Ballinger, DG, Cox, DR, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.
7. Lander, ES, Linton, LM, Birren, B, et al. (2001) Initial sequencing and analysis of the human genome. Nature **409**: 860–921.
8. Venter, JC, Adams, MD, Myers, EW, et al. (2001) The sequence of the human genome. *Science* **291:** 1304–1351.
9. Sachidanandam, R, Weissman, D, Schmidt, SC, et al. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.
10. Perkel, J. (2008) SNP genotyping: six technologies that keyed a revolution. *Nat Methods* **5:** 447–453.
11. Altshuler, D, Brooks, LD, Chakravarti, A, et al. (2005) A haplotype map of the human genome. *Nature* **437:** 1299–1320.
12. Hirschhorn, JN, Lohmueller, K, Byrne, E, et al. (2002) A comprehensive review of genetic association studies. *Genet Med* **4:** 45–61.
13. Lohmueller, KE, Pearce, CL, Pike, M, et al. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33:** 177–182.
14. Chanock, SJ, Manolio, T, Boehnke, M, et al. (2007) Replicating genotype-phenotype associations. *Nature* **447:** 655–660.
15. Saxena, R, Voight, BF, Lyssenko, V, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316:** 1331–1336.
16. Clayton, D, Leung, HT. (2007) An R package for analysis of whole-genome association studies. *Hum Hered* **64:** 45–51.
17. Abecasis, GR, Cardon, LR, Cookson, WO. (2000) A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66:** 279–292.
18. Laird, NM, Horvath, S, Xu, X. (2000) Implementing a unified approach to family-based tests of association. *Genet Epidemiol* **19** Suppl 1**:** S36–S42.
19. Lange, C, DeMeo, D, Silverman, EK, et al. (2004) PBAT: tools for family-based association studies. *Am J Hum Genet* **74:** 367–369.
20. Barrett, JC, Fry, B, Maller, J, et al. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.
21. Price, AL, Patterson, NJ, Plenge, RM, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909.
22. Li, Y, Abecasis, GR. (2006) Mach 1.0: rapid haplotype reconstruction and missing genotype inference. *Am J Hum Genet* S79: 2290.
23. The 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:** 661–678.
24. Ge, D, Zhang, K, Need, AC, et al. (2008) WGAViewer: software for genomic annotation of whole genome association studies. *Genome Res* **18:** 640–643.
25. Patterson, N, Hattangadi, N, Lane, B, et al. (2004) Methods for high-density admixture mapping of disease genes. *Am J Hum Genet* **74:** 979–1000.
26. Purcell, S, Cherny, SS, Sham, PC. (2003) Genetic power calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19:** 149–150.
27. Skol, AD, Scott, LJ, Abecasis, GR, et al. (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38:** 209–213.
28. Visscher, PM, Hill, WG, Wray, NR. (2008) Heritability in the genomics era–concepts and misconceptions. *Nat Rev Genet* **9:** 255–266.
29. Cupples, LA, Arruda, HT, Benjamin, EJ, et al. (2007) The Framingham Heart Study 100 K SNP genome-wide association study

resource: overview of 17 phenotype working group reports BMC. *Med Genet* 8 Suppl **1:** S1–S4.

30. Mitchell, AA, Cutler, DJ, Chakravarti, A. (2003) Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* **72:** 598–610.
31. Zondervan, KT, Cardon, LR. (2007) Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* **2:** 2492–2501.
32. de Bakker, PI, Yelensky, R, Pe'er, I, et al. (2005) Efficiency and power in genetic association studies. *Nat Genet* **37:** 1217–1223.
33. Pe'er, I, de Bakker, PI, Maller, J, et al. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat Genet* **38:** 663–667.
34. Barrett, JC, Cardon, LR. (2006) Evaluating coverage of genome-wide association studies. *Nat Genet* **38:** 659–662.
35. Bhangale, TR, Rieder, MJ, Nickerson, DA. (2008) Estimating coverage and power for genetic association studies using near-complete variation data. *Nat Genet* **40:** 841–843.
36. Rabbee, N, Speed, TP. (2006) A genotype calling algorithm for Affymetrix SNP arrays. *Bioinformatics* **22:** 7–12.
37. Wittke-Thompson, JK, Pluzhnikov, A, Cox, NJ. (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet* **76:** 967–986.
38. Cox, DG, Kraft, P. (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Hum Hered* **61:** 10–14.
39. Clayton, DG, Walker, NM, Smyth, DJ, et al. (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37:** 1243–1246.
40. Campbell, CD, Ogburn, EL, Lunetta, KL, et al. (2005) Demonstrating stratification in a European American population. *Nat Genet* **37:** 868–872.
41. Devlin, B, Roeder, K. (1999) Genomic control for association studies. *Biometrics* **55:** 997–1004.
42. Zheng, G, Freidlin, B, Gastwirth, JL. (2006) Robust genomic control for association studies. *Am J Hum Genet* **78:** 350–6.
43. Pritchard, JK, Stephens, M, Rosenberg, NA, et al. (2000) Association mapping in structured populations. *Am J Hum Genet* **67:** 170–181.
44. Purcell, S, Neale, B, Todd-Brown, K, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.
45. 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:** 1336–1341.
46. 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:** 1341–1345.
47. Lasky-Su, J, Lyon, HN, Emilsson, V, et al. (2008) On the replication of genetic associations: timing can be everything! *Am J Hum Genet* **82:** 849–858.
48. Balding, DJ. (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**: 781–791.
49. Liang, KY, Zeger, SL. (1986) Longitudinal data analysis using generalized estimating linear models. *Biometrika* **73:** 12–22.
50. Pilia, G, Chen, WM, Scuteri, A, et al. (2006) Heritability of cardiovascular and personality traits in 6,148 Sardinians. *PLoS Genet* **2:** e132.
51. Lowe, JK, Maller, JB, Pe'er l, et al. (2009) *PLoS Genet* **5:** e1000365.
52. Smith, JG, Lowe, JK, Kovvali, S, et al. (2009) *Heart Rhythm* **6:** 634–641.
53. Wacholder, S, Chanock, S, Garcia-Closas, M, et al. (2004) Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst* **96:** 434–442.
54. Dudbridge, F, Gusnanto, A. (2008) Estimation of significance thresholds for genome-wide association scans. *Genet Epidemiol* **32:** 227–234.
55. Pe'er, I, Yelensky, R, Altshuler, D, et al. (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32:** 381–385.
56. Lander, E, Kruglyak, L. (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* **11:** 241–247.
57. Kazeem, GR, Farrall, M. (2005) Integrating case-control and TDT studies. *Ann Hum Genet* **69:** 329–335.

58. Higgins, JP, Thompson, SG, Deeks, JJ, et al. (2003) Measuring inconsistency in meta-analyses. *BMJ* **327:** 557–560.
59. Willer, CJ, Sanna, S, Jackson, AU, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40:** 161–169.
60. 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:** 189–197.
61. Anderson, CA, Pettersson, FH, Barrett, JC, et al. (2008) Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am J Hum Genet* **83:** 112–119.
62. McCarthy, MI, Abecasis, GR, Cardon, LR, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* **9:** 356–369.
63. Lander, ES, Green, P, Abrahamson, J, et al. (1987) MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* **1:** 174–181.
64. Reich, D, Patterson, N, De Jager, PL, et al. (2005) A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility. *Nat Genet* **37:** 1113–1118.
65. Altmuller, J, Palmer, LJ, Fischer, G, et al. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* **69:** 936–950.
66. Altshuler, D, Daly, M. (2007) Guilt beyond a reasonable doubt. *Nat Genet* **39:** 813–815.
67. Benjamini, Y, Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc* **57:** 289–300.
68. Birney, E, Stamatoyannopoulos, JA, Dutta, A, et al. (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447:** 799–816.
69. Dixon, AL, Liang, L, Moffatt, MF, et al. (2007) A genome-wide association study of global gene expression. *Nat Genet* **39:** 1202–1207.
70. Stranger, BE, Nica, AC, Forrest, MS, et al. (2007) Population genomics of human gene expression. *Nat Genet* **39:** 1217–1224.
71. Kwan, T, Benovoy, D, Dias, C, et al. (2008) Genome-wide analysis of transcript isoform variation in humans. *Nat Genet* **40:** 225–231.
72. McCarroll, SA, Altshuler, DM. (2007) Copy-number variation and association studies of human disease. *Nat Genet* **39:** S37–S42.
73. Hirschhorn, JN, Daly, MJ. (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet* **6:** 95–108.
74. Wang, WY, Barratt, BJ, Clayton, DG, et al. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet* **6:** 109–118.
75. Pearson, TA, Manolio, TA. (2008) How to interpret a genome-wide association study. *JAMA* **299:** 1335–1344.
76. Kruglyak, L. (2008) The road to genome-wide association studies. *Nat Rev Genet* **9:** 314–318.