# Chapter 25

# Planning and Executing a Genome Wide Association Study (GWAS)

**Michèle M. Sale, Josyf C. Mychaleckyj, and Wei-Min Chen**

## Abstract

In recent years, genome-wide association approaches have proven a powerful and successful strategy to identify genetic contributors to complex traits, including a number of endocrine disorders. Their success has meant that genome wide association studies (GWAS) are fast becoming the default study design for discovery of new genetic variants that influence a clinical trait or phenotype. This chapter focuses on a number of key elements that require consideration for the successful conduct of a GWAS. Although many of the considerations are common to any genetic study, the greater cost, extreme multiple testing, and greater openness to data sharing require specific awareness and planning by investigators. In the section on designing a GWAS, we reflect on ethical considerations, study design, selection of phenotype/s, power considerations, sample tracking and storage issues, and genotyping product selection. During execution, important considerations include DNA quantity and preparation, genotyping methods, quality control checks of genotype data, *in silico* genotyping (imputation), tests of association, and replication of association signals. Although the field of human genetics is rapidly evolving, recent experiences can help guide an investigator in making practical and methodological choices that will eventually determine the overall quality of GWAS results. Given the investment to recruit patient populations or cohorts that are powered for a GWAS, and the still substantial costs associated with genotyping, it is helpful to be aware of these aspects to maximize the likelihood of success, especially where there is an opportunity for implementing them prospectively.

**Key words:** Linkage disequilibrium, haplotype, association, population, genetic, genome, genome wide, single nucleotide polymorphism.

## 1. Introduction

The sequencing of the human genome (1, 2), and SNP discovery and genotyping efforts (3) led to the discovery that the human genome is arranged in blocks of high linkage disequilibrium (LD), separated by hotspots of recombination (4). Resources from the

O.-K. Park-Sarge, T.E. Curry, Jr. (eds.), *Molecular Endocrinology,* Methods in Molecular Biology 590,
DOI 10.1007/978-1-60327-378-7_25, 

International HapMap Project (5–7) and affordable, accurate, high-throughput genotyping technologies have permitted analysis of the entire human genome using association methods, exploiting the fact that common variation in the human genome can be surveyed by genotyping only a fraction of the estimated 10–15 million single nucleotide polymorphisms (SNPs) that exist in the human population (8–10). This approach, using either tagging SNPs or protein coding non-synonymous SNPs, has revolutionized human genetics over the past 5 years. Recent analyses have identified loci for several endocrine measures and disorders, including type 2 diabetes (11, 12), type 1 diabetes and other autoimmune diseases (13), thyroid disease or thyroid measures

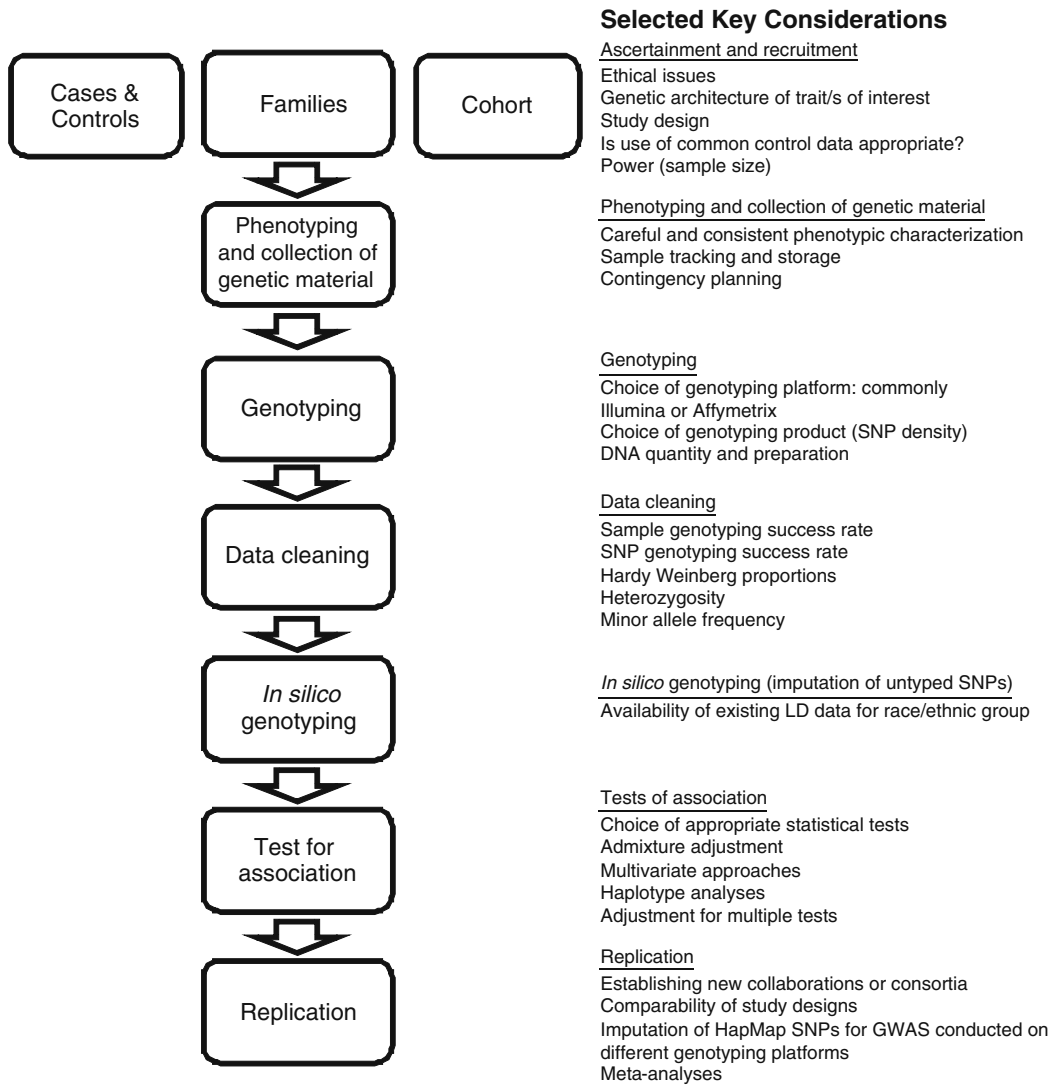


Fig. 25.1. Schematic of GWAS process.

(14, 15), bone mineral density and osteoporosis (16, 17), obesity (18–20), and adult height (21–24). In many cases, the newly identified associated genes and regions have provided insights on novel pathways and potential therapeutic targets (25).

This chapter focuses on a number of key elements to consider in the planning and successful conduct of a genome wide association study (GWAS), summarized in **Fig. 25.1**. Although this is an evolving and dynamic field of active research with many diverse opinions on approaches, some general recommendations can be made. Where possible, we point to in-depth reviews by other authors that provide more detailed explanation of these issues.

## 2. Planning a Genome Wide Association Study (GWAS)

### *2.1. Ethical Considerations*

The scope of existing consent documents is an important consideration before embarking on a GWAS. Many studies were designed and initiated prior to the availability of GWAS technology and did not have to confront the issues of individual identifiability and mandated sharing of data under national biobank policies. An example of such a policy is the current National Institutes of Health policy for awards that include a GWAS component (available at http://grants.nih.gov/grants/gwas/) and the deposition of de-identified data in the National Institutes of Health's Database of Genotype and Phenotype (dbGaP) (26). Recent policy changes reflect new concerns over potential identification of individual participants from summary data (27). Retrospective extension of existing clinical trial or observational studies to test genetic hypotheses raises major and profound issues of the applicability of an existing informed consent to the genetic study being considered. Consents may contain language that restricts data sharing to investigators of the trial or study only, or may preclude the use of the research data for studying diseases other the main study outcome. It is important to review consent wording in terms of participant intent and data sharing potential. For new studies, the possibility that de-identified data may be able to be matched against a second sample given by the same participant should be explicitly stated and explained during consent, since aggregated genetic data of this magnitude will undoubtedly constitute a unique genetic profile. Investigators are strongly urged to consult with local experts in ethics and genetics on their local research review board.

### *2.2. Study Design Issues*

Although family designs can be used for GWAS projects, and the resulting data can be analyzed under family-based tests of association (e.g., 28–29), many of the early successes with GWAS have

come from retrospective case-control study designs. Case-control designs offer advantages in relative efficiency of recruitment of independent sampling units, ease of analysis, and power for tests of common genetic variants that contribute only modest-to-low relative risk.

For a case-control design, matching of samples is critical to avoid biases that will inflate the overall type 1 error rate and lead to the identification of thousands of apparently significant SNPs. Matching can be performed at the individual level in a 1:1 or 1:*M* design (where *M* is an integer), or could be based on equalizing the frequency of known study subpopulations. For potential confounders for which matching may not be feasible, post-hoc adjustments can be performed during statistical analysis. It is critical that cases and controls are matched for ethnicity as closely as possible to avoid confounding and spurious associations. In this regard, GWA studies draw from the same design principles employed for any observational or clinical trial.

Several studies have demonstrated the utility of genotyping a common set of population controls for analyses of multiple traits. In particular, The Wellcome Trust Case-Control Consortium (WTCCC) (30) genotyped 500,000 SNPs in a common set of 3,000 controls drawn from the 1958 British birth cohort and the UK Blood Services collection, and have used these for a number of disease-based GWAS. Tests of association using a Cochran-Armitage additive trend statistic showed a high degree of concordance of the separately ascertained, but ethnically matched, UK controls. Similarly, use of previously genotyped population controls ascertained from an independent study can result in a considerable cost saving or increase in power (31), but the population must be substantially free of the disease or phenotype under investigation, and must be ethnically well-matched.

Many of the issues relating to study design, phenotype measures, and power are reviewed in (25, 32, 33). Considerations for designing a study, including the genetic architecture of complex traits, population stratification, and phenotype data are considered in detail by (34), while the role of family study designs in GWAS is reviewed in (35).

### 2.3. Selection of Phenotype/s

Often under-appreciated aspects of GWAS are the choice of phenotype definition and method of measurement of the primary phenotype and potential confounders. Large, well-designed prospective cohort studies are often advantageous since protocols are consistent across sites. However, even smaller single-site investigations need to consider the spectrum of phenotypic data that may be useful for GWAS analyses in order to provide insight into molecular mechanisms. Given the need for large-scale, often international, collaborative efforts to establish

replication or identify alleles of modest effect, it is helpful to develop standardized phenotypic protocols to facilitate comparable cross-study analyses and meta-analyses of data.

### 2.4. Power Considerations

The power to detect association is a function of the effect size, sample size (number of cases and controls, or families), and the tested association disease model. These factors are themselves influenced by the prevalence of the disease, disease allele frequency and the genotypic relative risk (GRR). For a typical study design that plans to genotype 1,000 cases and 1,000 controls for 300,000 markers, a disease-predisposing variant with GRR = 1.415 under a multiplicative model, with prevalence 0.1 and risk allele frequency 0.5 can be detected with 80% power. To reduce the high cost of genotyping, a two-stage design has been proposed where a proportion of samples are genotyped on every marker in stage 1, and a proportion of these markers are later followed up by genotyping them on the remaining samples in stage 2. For the above example, nearly the same power (77%) can be achieved with only 34% as many genotypes by using 30% of samples in stage 1 and 5% markers in stage 2. The software package CaTS (36) provides a convenient way for users to plan the sample size and power for their studies.

### 2.5. Sample Tracking and Storage

A known potential source of error for any GWAS is sample handling within the laboratory. A number of sample tracking and evaluation steps can be put in place to reduce the potential for sample mishandling or mislabeling. The National Cancer Institute's Office of Biorepositories and Biospecimen Research (http://biospecimens.cancer.gov/) is a useful resource for best practices and policies for biospecimen storage and tracking, and has also developed a suite of informatics tools available through the cancer Biomedical Informatics Grid (caBIG). At entry into the lab, a digital photographic record showing original sample sources (e.g., tubes) and label details is often useful. Allocation of a unique barcoded ID at the point of sample receipt enables tracking within the laboratory. Sample ID, receipt date, sample type, and storage location, together with any other available information, should be electronically logged into a secure database and, following DNA isolation, the quality and quantity of the genetic material should be assessed using standard methods (*see* **Section 3.1**). To enable back-referencing as the sample moves through sample processing stages (e.g., source sample, DNA isolation, aliquoting, dilution, plating, and genotyping), a number of alternate and complementary strategies can be implemented, depending on sample volume and financial resources. The simplest is to store a back-up sample of the original source material, such as an aliquot of frozen whole blood (or saliva, frozen tissue, etc.) or, for liquid samples, by spotting onto an FTA card (Whatman, Kent, UK) or equivalent product. These samples can be accessed later for DNA isolation

and genotyping, or direct PCR, if a sample handling mix-up is suspected. Another option is to conduct a simple PCR-based sex check immediately following DNA isolation, since this approach can often pick up mislabeling during recruitment, or wider sample handling problems during DNA isolation. Finally, generating a "mini-fingerprint" of highly polymorphic genetic variants (SNPs or microsatellites) on all incoming samples can serve as a more specific sample reference, but typically requires additional financial and personnel resources for sample processing and genotyping. The forensic community typically uses 13 short tandem repeat (STR) CODIS (Combined DNA Index System) markers that have been developed for utilization in forensic casework, although approximately 30 well-chosen SNPs would provide a similar information content.

For long-term storage, stock DNA samples can be placed in a −80°C ultra-low temperature freezers and should be stored in physically separate, duplicate locations on backup generator outlets (or in freezers with $CO_2$ backup). However, DNA samples that are accessed frequently can be stored at 4°C in the short-term to avoid multiple freeze-thaw cycles that can compromise DNA fragment length.

### 2.6. Genotyping Product Selection

The molecular methods described in this chapter utilize the Illumina Infinium assay (Illumina Inc., San Diego, CA) since we have most experience with this platform, however other genotyping platforms are available. The most commonly used alternative is the Affymetrix platform (Affymetrix Inc., Santa Clara, CA). Selection of the appropriate fixed content Illumina BeadChip is generally based on coverage in the population of interest and cost considerations. For example, the HumanCNV370-Quad BeadChip provides mean genomic coverage at $r^2 > 0.8$ for 0.87 of the genome, and, when combined with imputation methods, is generally adequate for populations of European ancestry where financial resources are limited (37). The higher SNP density of the Human1M-Duo BeadChip provides similar coverage (0.86 with $r^2 > 0.8$) in the HapMap Yoruba (YRI) population from Nigeria (http://www.illumina.com/pages.ilmn?ID=261).

## 3. Executing a GWAS

### 3.1. DNA Quantity and Preparation

Two key predictors of genotyping success are DNA quality and quantity. The 260/280 nm ratio, although a good measure of nucleic acid contamination of protein, is a poor measure of DNA contamination by protein (38). Visualization of DNA samples on gel to assess potential template degradation, and records of the sample storage and extraction methods, are likely to yield a more

accurate representation of overall quality. Although picogreen quantitation methods are recommended by Illumina Inc., we have had good success in GWAS assays using DNA samples quantitated using non-fluorescent methods, such as the NanoDrop-8000 spectrophotometer (NanoDrop Technologies; Wilmington, DE), provided input quantities exceed the specified minimum DNA quantity requirement. Illumina's fixed content GWAS products require a minimum input DNA of 400 ng for "Duo" products that process two samples per BeadChip, or 200 ng for "Quads" (four DNAs per BeadChip).

For a case-control study, case and control DNA samples should be intercalated in the wells of the genotyping daughter plates that will be used to prepare BeadChip assays. Case and control samples that have been separately ascertained, or collected and isolated, naturally lead to segregation of the case and control DNA samples into separate stock plates. While it is easiest to mirror this configuration of samples in daughter plates, plate-level biases in the genotyping assay or laboratory handling can lead to extreme plate-level effects that create spurious associations. If the samples are mixed on plates in approximately equal numbers, latent biases in the genotyping of an individual plate are less likely to generate spurious association results. From our experience, it is possible for a single plate to lead to spurious associations of more than 5 orders of magnitude in the measured *p*-values of association. Plate-level quality control checks of the samples in one plate compared to its sample complement should reveal plate(s) that appear to be outliers by SNP allele frequency or missing data.

### 3.2. Genotyping

Illumina's Infinium assay (39) is capable of multiplexing approximately 6,000 to 1 million SNPs/CNVs, either using fixed content products for GWAS, or customizable focused-content products (termed iSelect). At present, fixed content products for GWAS in humans range from approximately 370,000 to over 1 million markers per sample. Content is derived from HapMap data (6, 7), with a higher density of tagSNPs within 10 kb of a gene or in evolutionarily conserved regions.

In brief, Illumina's Infinium assay (39) consists of four modular components: (a) a single-tube whole-genome amplification step, (b) an array-based hybridization capture step, (c) an 'on array' enzymatic single base extension (SBE) step, and (d) an amplified-signal detection step. SBE uses a single 50 bp probe designed to hybridize adjacent to the SNP query site. After hybridization of target DNA to the BeadChip (a microelectromechanical systems (MEMS)-patterned substrate on silica slides), the SNP locus-specific primers, attached to 3-micron silica beads, are extended in the presence of hapten-labeled dideoxynucleotides. Biotin-labeled ddCTP and ddGTP, and 2,4-dinitrophenol (DNP)-labeled ddATP and ddUTP are efficiently incorporated

by polymerases and allow detection with a dual-color, orthogonal, multi-layer immune-histochemical sandwich assay. Biotin and DNP are simultaneously detected by staining with a combination of Alexa555-labeled streptavidin (SA) and Alexa647-labeled rabbit primary antibody against DNP, counterstaining with biotinylated anti-SA and DNP-labeled goat anti-rabbit secondary antibody. The signal is amplified by re-staining with Alexa555-SA/Alexa647-rabbit anti-DNP.

Physically separated pre- and post-PCR preparation areas are recommended to minimize possible cross-contamination of SBE primers from one assay to the next. It is advisable to use aerosol pipette tips, with separate boxes for pre- and post-PCR areas, and usual personal protective equipment such as latex gloves. Automation using robotics and incorporation of Laboratory Information Management Systems (LIMS), such as the Illumina's Infinium LIMS, can further reduce the possibility of error and contamination during processing of the Infinium assay.

Visualization of the resulting signal and decoding of SNP position is performed using a BeadArray Reader (Illumina Inc., San Diego, CA) and Illumina's proprietary data collection software. Data are initially analyzed using BeadStudio software (Illumina Inc., San Diego, CA), which automates clustering of genotypes and allele calling, as well as providing quality metrics to assist user inspection and removal of suspect data (*see* **Section 3.4**). This approach can be scaled to unlimited levels of multiplexing without compromising data quality, although in practical terms, the number of SNPs that can be assayed is limited by the number of probes on the array. Illumina has recently introduced a range of high-density (HD) BeadChips and has made minor adjustments to the protocol for this suite of products (40).

### 3.3. Quality Control Checks of Genotyping Data

Rigorous quality control is a crucial component of any GWAS since subtle biases in raw data can lead to hundreds or thousands of false positive results, confounding efforts to validate lead SNPs at the replication stage. Quality control steps to reject SNPs or samples are necessarily a trade-off between stringency to prevent type 1 error against loss of data, reducing power. The thresholds used in the individual steps reflect common values that are currently in use, but can be modified to be more or less tolerant of type 1 error. This decision will depend on study design, availability, and size of replication study samples, and willingness to include downstream manual steps to review cluster patterns of many SNP loci that appear to show significant association.

The first step involves use of vendor software to identify SNPs or samples that have obviously failed the assay or have generated significantly poorer quality data. Data checks from Infinium assays are initially conducted using Illumina's BeadStudio module. Given differences in allele frequencies, it is often advantageous to cluster

SNPs on the basis of individual race/ethnicity groups. Low-quality samples from the dataset can be identified and removed based on low signal intensities (less than 1000), low p10GC score (a quality metric, measuring distance from center of the cluster), and poor call rates (<95%). After removal of problem samples, all SNPs can be re-clustered and the call rates recalculated. SNPs with <95% call rate are classified as poor quality and removed. A large proportion of missing data indicates a non-robust SNP assay and is the best correlate of genotyping error or miscalling. We recommend the following pipeline of quality control checks:

1. Minor allele frequency (MAF): Retain only SNPs with MAF > 1%. Very low MAF SNPs are more susceptible to small biases in genotype calling algorithm.
2. Hardy Weinberg Equilibrium (HWE): Exclude SNPs with exact HWE test (41) jointly $p < 10^{-5}$. Deviations can indicate systematic genotype miscalling.
3. Cryptic duplicates: Test for outlier samples by examining the mean identity by state (IBS) between sample pairs by calculating the kappa coefficient. Pairs with extreme sharing suggest cryptic biological relationships or sample duplication.
4. Mean heterozygosity: Plots of heterozygosity distribution can reveal relative pairs (low) and contaminated samples (high).
5. X-linked heterozygosity: Plots of log odds ratio (sample is male/sample is female) can detect mis-specified study sex versus genetically inferred sex, indicating sample or data mix-up.

A potential problem for a GWAS is the presence of undetected population structure that may confound tests of association, leading to an increased rate of false positives or to false negative true associations (42). The effects of population structure increase with sample size, and for the size of study needed to detect typical genetic effects in common diseases, even the modest levels of population structure within population groups should not be ignored (43). A range of statistical issues for GWAS, including population substructure, are reviewed in (44). One approach is to use EIGENSTRAT (45, 46), which will detect outliers that are more than 6 standard deviations in any of 10 principal component (PC) dimensions by default, supplemented with a multivariate outlier detection algorithm. The optimal reduced principal components can be used as stratification adjustments in the generalized linear models for cross-sectional and longitudinal analysis (*see* **Section 3.5**). This approach has been shown to dramatically reduce the effects of population admixture, although we suggest carefully reviewing the principal components after computation and rationally choosing those to include as statistical model adjustments, to prevent over-adjustment.

Further genotype quality control measures are performed following initial analysis of association; this may involve visual

inspection of genotype cluster plots for the entire cohort for all SNPs deemed significant and all SNPs that show a distribution outside of HWE (at $P < 0.001$; this will include several thousand SNPs). This analysis is performed to ensure that there are no errors in automated calling and that any SNPs with potential errors can be discarded. Only if cluster plots reveal an obvious and correctible error is the SNP re-clustered and retained.

Quantile-quantile (QQ) plots are a particularly effective way to visually review the entire distribution of association or quality control statistics for all SNPs. These plots show the empirical distribution of statistics derived from the GWAS analysis plotted against the expected value for each ranked SNP under the global null hypothesis of no association or no significant test result for any SNP. A systematic difference in the empirical versus expected values across a fraction of the distribution may represent latent inflation of type 1 error through biases in the study populations, or may reflect true associations. Under a common variant complex model we expect to see true associated deviation from the expected null values for many SNPs in the more highly significant tail of the distribution.

As an illustrative example, genome-wide association scans were carried out using publicly available HapMap data (29, 47) to study the genetic basis of natural variation in gene expression. The data consist of gene-expression measurements (*CHI3L2* in this example) for 156 individuals in twenty 3-generation CEPH pedigrees, each with 12–17 individuals. Genotypes for 864,360 SNPs were generated for a subset of 90 individuals in these families in phase I of the International HapMap Project. Genotypes for 6,728 SNPs for the complete families, including 168 individuals, were also genotyped previously by the SNP Consortium. The GWA scan (*see* **Fig. 25.2**) maps gene *CHI3L2* to chromosome 1 with *p*-value $<10^{-9}$(29), and the *cis*-association has been confirmed by a functional assay analysis (47). The Q–Q plot for the GWA test statistics shows that overall the *p*-values are distributed uniformly between 0 and 1, and the log Q–log Q plot focuses attention on the tail of the distribution.

### 3.4. In Silico Genotyping

To improve genomic coverage of the selected marker panel, imputation methods have been implemented in software packages MACH (48) and IMPUTE (49) amongst others. Imputation improves the coverage of SNP panels as well as the power of a GWAS. Imputed SNPs are anticipated to lead to increased signal strength (lower *p*-values) near a signal from a true typed SNP, and also allow cross-genotyping platform analysis. The imputation method in MACH uses Markov models to identify stretches of shared chromosome between individuals, and then infer intervening genotypes by contrasting study samples with densely typed HapMap samples. Currently 2.5 million HapMap SNPs can be imputed for each individual, regardless of different genotyping

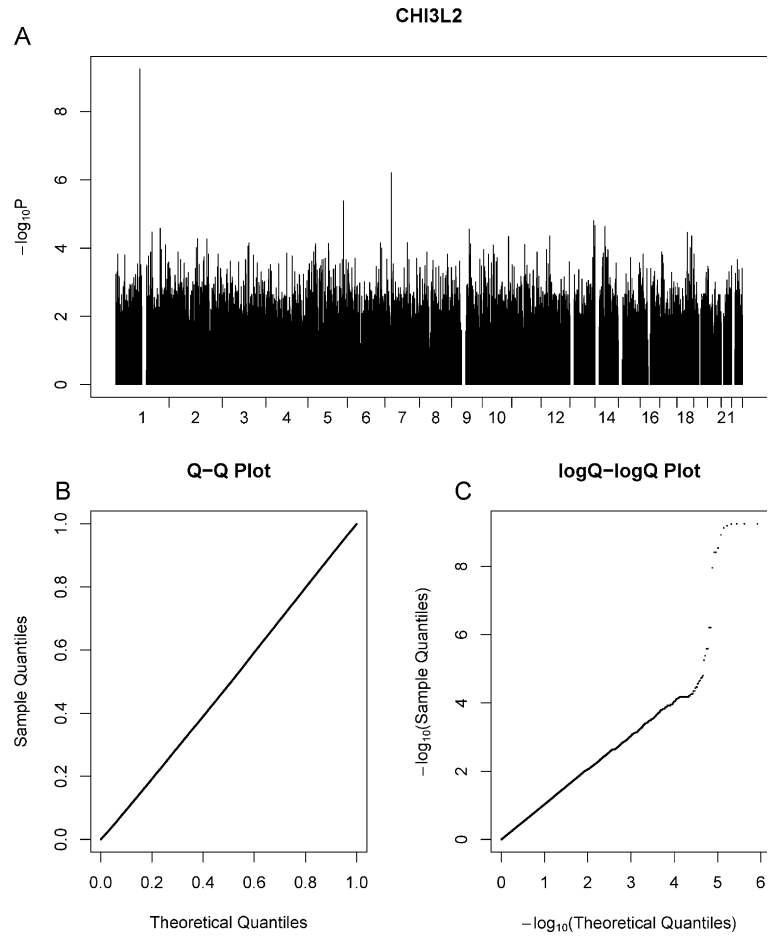


Fig. 25.2. Genome scan for *CHI3L2*expression levels. The gene maps to chromosome 1, and the association has also been confirmed by a functional analysis. (**A**) Genome scan using 90 individuals genotyped by the HapMap Consortium and 66 individuals with imputed genotypes. (**B**) Q–Q plot. (**C**) log Q–log Q plot.

platforms used. The imputation is accurate: in a few studies examined (50), focusing on top 95% top quality (measured by a correlation estimate between predicted and true genotypes) imputations, the error rate per SNP is about 1.1%. African samples were the most difficult to impute, with overall error rates ranging between 5.13% for the Yoruba and 11.86% for a sample from the San tribe when the HapMap YRI panel was used as a reference. In contrast, using the HapMap Centre d'Etude du Polymorphisme Human (CEPH) sample for European populations, and Chinese and Japanese HapMap samples for East Asian populations, resulted in overall error rates of <3.34 and 2.89% respectively (50). However, the accuracy can be further improved by tuning the quality metric threshold. It is prudent to confirm associated imputed genotyped by subsequent direct genotyping. For example, in a GWAS of

fasting glucose, the imputed SNP with the strongest association in gene G6PC2 was later genotyped and the discrepancy rate per allele between the imputed and typed genotypes was 1.4% (51).

### 3.5. Tests of Association

Although a variety of approaches can be used to analyze a GWAS, in this section we suggest some widely used applications, as well as a method uniquely capable of handling multivariate data. For single SNP analyses in case-control datasets, PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) (52) has been developed specifically for the analysis of GWAS data. For example, point tests of association can be estimated for each SNP using standard allelic and inheritance model association tests, or the Cochran-Armitage test for trend. For both discrete and quantitative traits, univariate and multivariable analyses can be performed to examine the contribution at each SNP to the specific trait of interest, depending on the hypothesis. The baseline linear regression framework allows for adjustment for environmental and other factors known or suspected to be confounding variables (e.g., age, sex, EIGENSTRAT principal components), as well as gene-by-gene and gene-by-environment interactions. Under a model of stratification, Cochran-Mantel-Haenzsel tests the conditional independence of the case-control disease status.

A multivariate trait GWA algorithm has been implemented the software package Ghost (Chen WM, personal communication, http://people.virginia.edu/~wc9c/ghost/). This implementation can help systematically identify genetic variants that are responsible for multiple traits.

For longitudinal data, the generalized estimating equation (GEE) method, using a sandwich estimator of the variance under an exchangeable correlation and allowing adjustment for covariates as well as gene-by-gene and gene-by-environment interactions, can take into account correlations of repeated measures, and association results will be less sensitive to the trait distribution (and thus more robust). Note this GEE based GWAS does not add computational complexity, and permutation testing can be easily followed to adjust for multiple testing.

Although it is possible under disease models of SNP-SNP interaction that haplotype tests could have greater power to detect association than single SNP tests, genome wide haplotype testing increases the total number of GWAS hypothesis tests conducted, and leads to reduced power resulting from the necessary increased stringency threshold for increased multiple testing. The usual approach is to perform single SNP tests first, utilizing the high accuracy of imputation methods to effectively increase marker density, and follow-up with targeted haplotype analyses where an associated SNP is contained within a haplotype block (4). Few, if any, current studies have power to detect and fully replicate genetic predisposition arising from gene–gene or SNP–SNP interactions.

To assess whether SNPs or haplotypes are associated with clinical outcomes, the Kaplan–Meier method of survival analysis can be used to estimate the survival functions for subjects with different genotypes in the follow up period, and a logrank test performed to compare the survival distributions. Cox's proportional hazards model can also be used by treating the genotype as a risk factor.

A number of analytical approaches for analyses of GWAS data are reviewed in (53).

### 3.6. Replication

Since the discovery stage of most GWAS designs is underpowered to detect the modest effects observed for the majority of complex diseases, such studies are anticipated to generate a substantial number of false positive results (type 1 errors) (54). Additionally, the initial effect estimates are likely to be inflated due to the phenomenon known as the "winner's curse" (55, 56). Replication in an independent population therefore remains a critical step in a GWAS to confirm initial results (57). Several different strategies have been employed, including the two-stage design mentioned previously (*see* **Section 2.4**). However, due to the large sample sizes required, large-scale meta-analyses across several independent studies have been a mainstay of GWAS reports. These studies frequently combine heterogeneous study designs, ascertainment and recruitment criteria, genotyping platforms, and QC metrics. A set of useful guidelines for imputation and meta-analyses have been developed by (58), and many of the relative merits, caveats, statistical approaches, and diagnostic tests for meta-analyses are reviewed in (59). It should be noted that lack of replication may reflect any one of a number of possible scenarios, including appropriate refutation of a false lead, false non-replication, or true genetic heterogeneity across studies (60).

## 4. Summary and Future Directions

Recent successes using GWAS approaches have generated considerable enthusiasm about the utility of this approach to identify variants that contribute to human variation and disease susceptibility. In a relatively short space of time, considerable advances have already been made in genotyping efficiency and cost, imputation approaches, and analytical methods. In future, further understanding of the roles of epistasis (gene–gene interactions), gene–environment interactions, copy number variants, and epigenetic phenomena are anticipated to provide additional insights into our understanding of complex human disorders.

## Acknowledgments

We wish to thank Andrew Singleton, Ph.D., National Institute on Aging, Kathleen H. Day, University of Virginia and Fang-Chi Hsu, Ph.D., Wake Forest University School of Medicine, for helpful discussion and comments.

### References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W et al.: Initial sequencing and analysis of the human genome. *Nature* 2001, 409(6822):860–921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA et al.: The sequence of the human genome. *Science* 2001, 291(5507):1304–1351.
3. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL et al.: A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001, 409(6822):928–933.
4. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M et al.: The structure of haplotype blocks in the human genome. *Science* 2002, 296(5576):2225–2229.
5. Olivier M: A haplotype map of the human genome. *Physiol Genomics* 2003, 13(1):3–9.
6. The International HapMap Consortium: A haplotype map of the human genome. *Nature* 2005, 437(7063):1299–1320.
7. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM et al.: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007, 449(7164):851–861.
8. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: Efficiency and power in genetic association studies. *Nat Genet* 2005, 37(11):1217–1223.
9. Gu CC, Yu K, Ketkar S, Templeton AR, Rao DC: On transferability of genome-wide tagSNPs. *Genet Epidemiol* 2008, 32(2): 89–97.
10. Gu CC, Yu K, Rao DC: Characterization of LD structures and the utility of HapMap in genetic association studies. *Adv Genet* 2008, 60:407–435.
11. Frayling TM: A new era in finding Type 2 diabetes genes-the unusual suspects. *Diabet Med* 2007, 24(7):696–701.
12. Lindgren CM, McCarthy MI: Mechanisms of disease: genetic insights into the etiology of type 2 diabetes and obesity. *Nat Clin Pract Endocrinol Metab* 2008, 4(3): 156–163.
13. Duffy DL: Genetic determinants of diabetes are similarly associated with other immune-mediated diseases. *Curr Opin Allergy Clin Immunol* 2007, 7(6):468–474.
14. Hwang SJ, Yang Q, Meigs JB, Pearce EN, Fox CS: A genome-wide association for kidney function and endocrine-related traits in the NHLBI's Framingham Heart Study. *BMC Med Genet* 2007, 8 Suppl 1:S10.
15. Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F et al.: Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* 2007, 39(7):857–864.
16. Richards JB, Rivadeneira F, Inouye M, Pastinen TM, Soranzo N, Wilson SG, Andrew T, Falchi M, Gwilliam R, Ahmadi KR et al.: Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* 2008, 371(9623): 1505–1512.
17. Kiel DP, Demissie S, Dupuis J, Lunetta KL, Murabito JM, Karasik D: Genome-wide association with bone mass and geometry in the Framingham Heart Study. *BMC Med Genet* 2007, 8 Suppl 1:S14.
18. Fox CS, Heard-Costa N, Cupples LA, Dupuis J, Vasan RS, Atwood LD: Genome-wide association to body mass index and waist circumference: the Framingham Heart Study 100 K project. *BMC Med Genet* 2007, 8 Suppl 1:S18.
19. Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, Perry JR, Elliott KS, Lango H, Rayner NW et al.: A common variant in the FTO gene is

associated with body mass index and predisposes to childhood and adult obesity. *Science* 2007, 316(5826):889–894.

20. Scuteri A, Sanna S, Chen WM, Uda M, Albai G, Strait J, Najjar S, Nagaraja R, Orru M, Usala G et al.: Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS Genet* 2007, 3(7):e115.
21. Weedon MN, Lettre G, Freathy RM, Lindgren CM, Voight BF, Perry JR, Elliott KS, Hackett R, Guiducci C, Shields B et al.: A common variant of HMGA2 is associated with adult and childhood height in the general population. *Nat Genet* 2007, 39(10): 1245–1250.
22. Lettre G, Jackson AU, Gieger C, Schumacher FR, Berndt SI, Sanna S, Eyheramendy S, Voight BF, Butler JL, Guiducci C et al.: Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat Genet* 2008, 40(5): 584–591.
23. Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, Bonnycastle LL, Shen H, Timpson N, Lettre G, Usala G et al.: Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* 2008, 40(2):198–203.
24. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, Mangino M, Freathy RM, Perry JR, Stevens S, Hall AS et al.: Genome-wide association analysis identifies 20 loci that influence adult height. *Nat Genet* 2008, 40(5):575–583.
25. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008, 9(5):356–369.
26. Mailman MD, Feolo M, Jin Y, Kimura M, Tryka K, Bagoutdinov R, Hao L, Kiang A, Paschall J, Phan L et al.: The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007, 39(10):1181–1186.
27. Homer N, Szelinger S, Redman M, Duggan D, Tembe W, Muehling J, Pearson JV, Stephan DA, Nelson SF, Craig DW: Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet* 2008, 4(8):e1000167.
28. Martin ER, Monks SA, Warren LL, Kaplan NL: A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 2000, 67(1): 146–154.
29. Chen WM, Abecasis GR: Family-based association tests for genomewide association scans. *Am J Hum Genet* 2007, 81(5): 913–926.
30. Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007, 447(7145): 661–678.
31. Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC et al.: Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat Genet* 2008.
32. Zondervan KT, Cardon LR: Designing candidate gene and genome-wide case-control association studies. *Nat Protoc* 2007, 2(10):2492–2501.
33. Amos CI: Successful design and conduct of genome-wide association studies. *Hum Mol Genet* 2007, 16 Spec No. 2:R220–225.
34. Kraft P, Cox DG: Study designs for genome-wide association studies. *Adv Genet* 2008, 60:465–504.
35. Cupples LA: Family study designs in the age of genome-wide association studies: experience from the Framingham Heart Study. *Curr Opin Lipidol* 2008, 19(2):144–150.
36. Skol AD, Scott LJ, Abecasis GR, Boehnke M: Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* 2006, 38(2):209–213.
37. Anderson CA, Pettersson FH, Barrett JC, Zhuang JJ, Ragoussis J, Cardon LR, Morris AP: Evaluating the effects of imputation on the power, coverage, and cost efficiency of genome-wide SNP platforms. *Am J Hum Genet* 2008, 83(1):112–119.
38. Glasel JA: Validity of nucleic acid purities monitored by 260 nm/280 nm absorbance ratios. *Biotechniques* 1995, 18(1):62–63.
39. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL: Whole-genome genotyping with the single-base extension assay. *Nat Methods* 2006, 3(1):31–33.
40. Illumina Inc.: Infinium HD Assay Super, Manual – Experienced User Card. In.: Part # 11294825.
41. Wigginton JE, Cutler DJ, Abecasis GR: A note on exact tests of Hardy-Weinberg equilibrium. *Am J Hum Genet* 2005, 76(5):887–893.
42. Pritchard JK, Donnelly P: Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 2001, 60(3): 227–237.

43. Marchini J, Cardon LR, Phillips MS, Donnelly P: The effects of human population structure on large genetic association studies. *Nat Genet* 2004, 36(5):512–517.
44. Teo YY: Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Curr Opin Lipidol* 2008, 19(2):133–143.
45. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006, 38(8):904–909.
46. Li Q, Yu K: Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol* 2008, 32(3):215–226.
47. Cheung VG, Spielman RS, Ewens KG, Weber TM, Morley M, Burdick JT: Mapping determinants of human gene expression by regional and genome-wide association. *Nature* 2005, 437(7063): 1365–1369.
48. Li Y, Abecasis GR: Mach 1.0: Rapid haplotype reconstruction and missing genotype inference. *American Journal of Human Genetics* 2006, S79:2290.
49. Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007, 39(7):906–913.
50. Huang L, Li Y, Singleton AB, Hardy JA, AbeCasis G, Rosenberg NA, Scheet P: Genotype-imputation accuracy across worldwide human populations. *American Journal of Human Genetics* 2009, 84(2):230–250.
51. Chen WM, Erdos MR, Jackson AU, Saxena R, Sanna S, Silver KD, Timpson NJ, Hansen T, Orru M, Grazia Piras M et al.: Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels. *J Clin Invest* 2008, 118(7):2620–2628.
52. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007, 81(3):559–575.
53. Langefeld CD, Fingerlin TE: Association methods in human genetics. *Methods Mol Biol* 2007, 404:431–460.
54. Senn S: Transposed conditionals, shrinkage, and direct and indirect unbiasedness. *Epidemiology* 2008, 19(5):652–654; discussion 657–658.
55. Ioannidis JP: Why most discovered true associations are inflated. *Epidemiology* 2008, 19(5):640–648.
56. Kraft P: Curses—winner's and otherwise—in genetic epidemiology. *Epidemiology* 2008, 19(5):649–651; discussion 657-648.
57. Willett WC: The search for truth must go beyond statistics. *Epidemiology* 2008, 19(5): 655–656.
58. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF: Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* 2008, 17(R2):R122-128.
59. Kavvoura FK, Ioannidis JP: Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum Genet* 2008, 123(1):1–14.
60. Ioannidis JP: Non-replication and inconsistency in the genome-wide association setting. *Hum Heredity* 2007, 64(4): 203–213.