Published in final edited form as:
Curr Protoc Hum Genet. 2011 January ; CHAPTER: Unit1.19. doi:10.1002/0471142905.hg0119s68.

Quality Control Procedures for Genome Wide Association Studies

Stephen Turner[1], **Loren L. Armstrong**[2], **Yuki Bradford**[1], **Christopher S. Carlson**[3], **Dana C. Crawford**[1], **Andrew T. Crenshaw**[4], **Mariza de Andrade**[5], **Kimberly F. Doheny**[6], **Jonathan L. Haines**[1], **Geoffrey Hayes**[2], **Gail Jarvik**[7], **Lan Jiang**[1], **Iftikhar J. Kullo**[8], **Rongling Li**[9], **Hua Ling**[6], **Teri A. Manolio**[9], **Martha Matsumoto**[5], **Catherine A. McCarty**[10], **Andrew N. McDavid**[3], **Daniel B. Mirel**[4], **Justin E. Paschall**[11], **Elizabeth W. Pugh**[6], **Luke V. Rasmussen**[10], **Russell A. Wilke**[12], **Rebecca L. Zuvich**[1], and **Marylyn D. Ritchie**[1]

[1]Center for Human Genetics Research, Department of Molecular Physiology & Biophysics, Vanderbilt University, Nashville, TN, USA.

[2]Division of Endocrinology, Metabolism, and Molecular Medicine, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA.

[3]Cancer Prevention, Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

[4]Genetic Analysis Platform and Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA.

[5]Division of Biostatistics and Informatics, Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, USA.

[6]Center for Inherited Disease Research, Johns Hopkins University, Baltimore, MD, USA.

[7]Department of Genome Sciences, University of Washington, Seattle, WA, USA, USA.

[8]Division of Cardiovascular Diseases, Department of Medicine, Mayo Clinic, Rochester, MN, USA.

[9]Office of Population Genomics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA.

[10]Biomedical Informatics Research Center, Marshfield Clinic Research Foundation, Marshfield, WI, USA.

[11]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA.

[12]Division of Clinical Pharmacology, Department of Medicine, Vanderbilt University, Nashville, TN, USA.

Abstract

Genome-wide association studies (GWAS) are being conducted at an unprecedented rate in population-based cohorts and have increased our understanding of the pathophysiology of complex disease. The recent application of GWAS to clinic-based cohorts has also yielded genetic predictors of clinical outcomes. Regardless of context, the practical utility of this information will

Corresponding Author Contact Information: Marylyn D. Ritchie, Ph.D. Associate Professor of Molecular Physiology & Biophysics, Division of Human Genomics Director, Program in Computational Genomics Center for Human Genetics Research Vanderbilt University School of Medicine 519 Light Hall Nashville, TN 37232-0700.

ultimately depend upon the quality of the original data. Quality control (QC) procedures for GWAS are computationally intensive, operationally challenging, and constantly evolving. With each new dataset, new realities are discovered about GWAS data and best practices continue to be developed. The Genomics Workgroup of the National Human Genome Research Institute (NHGRI) funded electronic Medical Records and Genomics (eMERGE) network has invested considerable effort in developing strategies for QC of these data. The lessons learned by this group will be valuable for other investigators dealing with large scale genomic datasets. Here we enumerate some of the challenges in QC of GWAS data and describe the approaches that the eMERGE network is using for quality assurance in GWAS data, thereby minimizing potential bias and error in GWAS results. In this protocol we discuss common issues associated with QC of GWAS data, including data file formats, software packages for data manipulation and analysis, sex chromosome anomalies, sample identity, sample relatedness, population substructure, batch effects, and marker quality. We propose best practices and discuss areas of ongoing and future research.

INTRODUCTION

Genome-wide association studies (GWAS) are commonly used to identify common single nucleotide polymorphisms (SNPs) that influence human traits. GWAS have been conducted at increasing frequency using case-control, population-based prospective, and cross-sectional study designs [1-6]. More recently, GWAS are being conducted in cohorts that are clinic-based [7-10]. As a result, GWAS may soon move the field of genomics into clinical practice.

Whether the goal is to identify predictors of outcomes or to discover new biology underlying a trait of interest, the capability of GWAS to identify true genetic associations depends upon the overall quality of the data. Even simple statistical tests of association are compromised in the context of genome-wide SNP data that have not been properly cleaned, potentially leading to false-negatives and false-positive associations. Additionally, problems with the overall data quality will likely affect downstream analyses and studies beyond the initial GWAS. For example, the National Human Genome Research Institute (NHGRI) actively maintains an online catalog of GWAS results and associated publications [6], which stimulates downstream studies of replication and characterization in independent populations. Compromised data quality in the discovery phase may lead to false positive results that are carried forward into replication studies at great cost both in time and expense. Also, the National Institutes of Health (NIH) now mandates that secure, encrypted copies of primary GWAS data funded by NIH be made publicly available (with controlled access) for secondary analyses. These accessible datasets are maintained by the National Center for Biotechnology Information (NCBI) in the database of Genotypes and Phenotypes (dbGaP). dbGaP provides both open and controlled access, which allow for both broad release of non-sensitive information, and restricted access to datasets involving genomic data and phenotypic information, respectively [11]. Data access through dbGaP is commonly used for replication and meta-analysis, both of which will be compromised by poor quality data.

Genotyping technology and allele calling algorithms continue to improve and quality-improvement strategies continue to ensure that only reliable, rigorously scrutinized markers and samples are used for analysis. Reconciling genetic data with clinical and self-reported data (e.g., sex or familial relationships) can potentially identify sample identity problems caused by sample handling mishaps. Batch effects, population stratification, and sample relatedness can confound genetic association analyses and can lead to excessive type I and type II errors. Here we discuss methods that can be used to detect and account for various

data quality issues to better ensure the integrity of the primary GWAS as well as its downstream applications.

The eMERGE (electronic MEdical Records and GEnomics) Network is an NHGRI-supported consortium of five institutions charged with exploring the utility of DNA repositories coupled to Electronic Medical Record (EMR) systems for advancing discovery in genome science [12]. Genome-wide genotyping has been performed on ~17,000 samples across the eMERGE network at the Broad Institute and at the Center for Inherited Disease Research (CIDR) using the Illumina 660W-Quad or 1M-Duo Beadchips. Each study site is conducting a GWAS, in addition to a number of cross-network analyses. These studies adhere to NIH's data sharing policies, and all data generated in this study will be available on dbGaP [11]. Due to the complexity involved in a single site GWAS, in addition to the combining of data and results across study sites, it became clear that a unified QC pipeline was imperative.

Others have discussed quality control procedures for genotypic data [13-16]. The goal of this manuscript is a tutorial to instruct investigators on QC procedures that should be performed prior to GWAS data analysis. The procedures discussed here were developed by the genomics group of the eMERGE network, where phenotyping and other sample information is obtained through sophisticated mining of the EMR. This protocol can be applied to many GWAS studies, regardless of phenotyping strategy. Given that most of the genotyping data available for GWAS is currently SNP-based, we will limit our discussion to these biallelic markers, and QC procedures for CNV analysis will not be discussed here. Figure 1 shows a flowchart overview of the entire QC process, where each step is discussed in detail in the following sections.

GWAS DATA FORMAT

Regardless of the underlying study design (such as family-based or population-based), the most commonly used format for genetic data is the linkage, or pedigree file format (pedfile). This file contains one individual per row, where the first six columns are identifying information (family ID, individual ID, father ID, mother ID, sex, phenotype), and the remaining columns are genotypes (2 columns per genotype; one for each allele). The genotype column-pairs correspond to an ordered set of SNP markers present in an associated file (.map or .bim). Additional phenotypes can also be stored in separate files consisting of family ID, individual ID, then extra columns representing additional phenotypes. There are several variations on pedfile format, including transposed (long) formats (tped), and compressed (binary) formats. Descriptions of these file formats can be found on the PLINK homepage (Table 1). PLINK is a freely available, open source, cross platform application for QC and analysis of GWAS data [17]. We used PLINK for implementing most of the eMERGE network's QC pipeline.

An important issue when creating a pedfile for QC analysis is the choice of strand orientation to use for allele calls (i.e., forward or reverse complement). While forward strand is a commonly used allele coding scheme, Illumina has developed a consistent and simple method to ensure uniformity in genotype call reporting that uses the polymorphism itself and the contextual surrounding sequence ("TOP/BOT" strand and "A/B" allele coding) [18]. Since 2005, the database of genetic variation (dbSNP) [19] has used this designation for all SNP entries. We used "TOP/BOT strand orientation for eMERGE. Choice of strand orientation might depend on the strand orientation of other data used in a combined analysis or of a reference set used for imputation. The goal is to ensure uniformity in genotype call reporting that is critically important in downstream analyses, reporting, and annotation.

SAMPLE QUALITY

Sex inconsistencies and chromosomal anomalies

One of the first procedures that should be implemented in any GWAS QC protocol is checking for potential sample identity problems that typically result from sample handling errors. One of the easiest ways to discover potential sample handling issues that result in mix-ups is by checking the reported sex of each individual against that predicted by the genetic data. The -- check-sex option in PLINK uses X chromosome heterozygosity rates to determine sex empirically, then reports individuals for whom the sex recorded in the pedfile does not match the predicted sex based on genetic data (example output and explanation shown in Table 2). If discrepancies are found (e.g. an individual is recorded being female but appears homozygous for every X chromosome marker), the EMR or any available study questionnaires should be reviewed to make a determination whether there was a sample handling mistake that caused a sample mix-up. Checking X chromosome heterozygosity may also reveal sex chromosome anomalies such as Turner syndrome (females having karyotype XO), Kleinfelter syndrome (males having karyotype XXY), mosaic individuals (XX/XO, XX/XXY), or females with large stretches of loss-of-heterozygosity on the X chromosome who are otherwise phenotypically normal.

X chromosome heterozygosity is a fairly sensitive heuristic to detect sample swaps, but not very specific. A variety of factors besides a crude sample mix-up will affect heterozygosity. Furthermore, if the goal is to enumerate as many samples with atypical sex karyotypes as possible, then X heterozygosity alone will not detect abnormalities such as triple X or XYY or homozygous X Kleinfelter syndrome. Examining the intensity of probe binding on the sex chromosomes will better resolve these cases. Illumina calls this intensity LogR ratio. On Affymetrix systems, it is simply known as probe intensity. These metrics, once suitably normalized, are roughly linear in copy number. Because there are tens of thousands of loci on the X chromosome on modern platforms, it is appropriate to examine a subsample of markers, and then take a measure of central tendency of each sample such as the median or mean intensity. The intensity plot provides a visualization of the intensity of X and Y probes (Figure 2). It is expected that females should have low Y intensity and high X intensity (bottom right corner), and the males show have similar level of X and Y intensities (top left corner). also In our dataset of 17,000 individuals two individuals were observed with mislabeled sex and XX individuals with XXY. Structural chromosomal variation can be identified using intensity only probes to calculate loss of heterozygosity and abnormal copy numbers using B allele frequency and Log R Ratio plots (Figure 3). The B allele frequency plot is the amount of B allele observed in a probe that should concentrate at zero for zero copy, at 0.5 for one copy and at 1 for two copies. Log R ratio is the ratio of a particular sample overall all samples. It is expected to concentrate at zero; sometimes an upward or downward bump is observed meaning amplification or deletion, respectively. These two plots can be obtained using Illumina BeadStudio/GenomeStudio.

Depending on the aims of the study, often these individuals are not eliminated from the study due to sex chromosome anomalies alone. Even in carefully collected samples, the numbers of samples with discrepant self-reported sex having a normal karyotype is appreciable, and sample processing pipelines need to have these checks in place to detect such potential sample swaps. While it may be possible to go back to the original data to reconcile chromosomal anomalies found using genetic data, researchers need to be aware of any ethical issues that may arise concerning return of results, and these issues should be considered and resolved prior to revisiting the EMR.

Sample relatedness

Another way to simultaneously examine both sample identity and pedigree integrity is by reconciling genomic data with self-reported relationships between individuals (if available). Although this has consequences that impact which analytical approaches are appropriate for the downstream association study, having related samples in the dataset makes it possible to further investigate potential DNA sample mix-ups. Using dense marker data obtained in GWAS, it is easy to compute pairwise kinship estimates between every individual in the study using the -- genome option in PLINK. This procedure need not be performed on the entire GWAS using – dataset only 100,000 markers will also yield stable estimates of kinship coefficients. In addition to reporting the relationship type as reported using pedigree data (e.g. siblings, parent-child, unrelated), this procedure will also calculate the proportion of loci where two individuals share zero, one, or two alleles identical by descent (IBD). Individuals sharing two alleles IBD at every locus are either monozygotic twins, or the pair is actually a single sample processed twice. Individuals sharing zero alleles IBD at every locus are unrelated. Individuals sharing one allele IBD at every locus are parent-child pairs. On average, siblings share zero, one, and two alleles IBD at 25%, 50%, and 25% of the genome, respectively. Using these data, the proportion of loci sharing one allele IBD (the Z1 column) by the proportion of loci where individuals share zero alleles IBD (column Z0) can be plotted and points color coded by the relationship type. For clarity, this plot can be restricted to points where the overall kinship coefficient is ≥ 0.05, as most of the individuals where kinship ≤ 0.05 will be unrelated. This will produce a plot as shown in Figure 4. Detailed instructions on producing this graphic using R [20] can be found online [21]. If it is believed that pedigree records obtained through the original data are accurate, then a point out of place (e.g., points colored as unrelated showing up where most of the parent-offspring pairs cluster) would be indicative of either nonpaternity, adoption, sample mix-up, or duplicate processing of a single individual. Further investigation using the original data can be used to attempt to identify the problem. It is also worth noting in studies where datasets from multiple sites are combined that it is possible that the same participant is present in more than one study. These two data points would appear genetically identical across sites.

In addition to potentially discovering sample handling issues, visualizing sample relatedness as shown in Figure 4 also reveals any cryptic relatedness that may be present in the study sample. Figure 4 shows that many individuals who indicated that they were unrelated (black points) or distantly related (blue points) line up along the diagonal in this plot. These individuals represent second, third, fourth, and fifth degree relatives. If treated as independent samples in the downstream analyses, having many related samples in the dataset would result in increased type I and type II errors, thus analytical methods such as mixed model regression [22] must be used in place of simple linear or logistic regression. Figure 5 shows another way to visualize the degree of relatedness by plotting a histogram of the distribution of kinship coefficients over 0.05 between all pairs of individuals in the dataset.

Population substructure

Population stratification occurs when the study samples comprise multiple groups of individuals who differ systematically in both genetic ancestry and the phenotype under investigation. Spurious apparent associations would be due to differences in ancestry rather than true association of alleles to disease [23]. Thus it is critical to check for population stratification within the study samples and leverage this information to inform the downstream analyses.

One strategy for avoiding bias induced by population stratification is to ensure that study samples are drawn from a relatively homogenous population. One of the sites in the

eMERGE network represents such a sample, as over 98% of the study sample self-reported "Caucasian," on a study questionnaire. This percentage is consistent with data from the 2000 Census [24], and self-report often shows very high correspondence with genetically inferred ancestry [25]. Some clinics record ethnicity via observer-report (typically a clerk or nurse's aide). Even in this settings, observer-reported ancestry closely matches genetically inferred ancestry, especially for populations of European descent [26]. However, population-based diverse samples are often desirable for genetic association studies focused on characterizing previous GWAS or candidate gene discoveries made in one population [27]. Further, combining samples from multiple sites for a joint analysis may result in population stratification in the combined sample, if both allele frequencies and outcomes differ between sites.

Statistical methodology has been developed and implemented into software to aid in detecting and adjusting for population stratification in GWAS. Genomic control [28,29], aims to control for population stratification by first estimating an inflation factor, then adjusting all of the test statistics downward by this factor. Several variations on genomic control have been developed, and a recent review and critical evaluation of genomic control methods [30] recommended genomic control F (GCF) [31] as the most appropriate variation. GCF does not assume the inflation factor is measured without error, and refines this factor accordingly. Structured association [32], implemented in the STRUCTURE software [33] uses genotype data to infer population structure and subsequently performs tests of association within each inferred subpopulation. STRUCTURE may also be used to identify individual samples that do not cluster with the majority of the samples. These samples can then be eliminated from the analysis.

Because the risk of confounding by population stratification may increase with sample size (i.e., confounded results become more significant with larger samples) [34], and because large sample GWAS are becoming increasingly common, another method has been developed that utilizes large samples and thousands of markers throughout the genome to adjust for population structure. Eigenstrat analysis [35,36] uses principal components analysis to explicitly detect and adjust for population stratification on a genome-wide scale in large sample sizes in a computationally efficient manner. This method may be preferred over a stratified analysis because the combined sample often yields more powerful statistical tests, even after adjusting for significant eigenvectors [37]. Eigensoft is freely available open-source software for conducting Eigenstrat analyses, available online (Table 1). Running Eigensoft requires dense genotyping coverage. We recommend using all the default options, including 100,000 randomly chosen high-quality markers. There are several SNPs in the HLA region on chromosome 6, in the lactase locus on chromosome 2, and in the inversion regions on 8p23 and 17q21.31 common in populations of European ancestry [38] that are sources of stratification that will often appear in the top principal components. While one may exclude these SNPs from such an analysis, it is unknown if any similar inversions exist at appreciable frequency in non-European population. Thus it may be preferable to detect and correctly interpret the analysis including these regions rather than avoiding specific regions. The Eigensoft analysis will result in the computation of 10 principal components. If any of these eigenvectors are significantly associated with the phenotype under study, it is recommended that these eigenvectors be adjusted for in any downstream analysis to correct for any bias due to population stratification. Alternatively, if it is expected that only a very small number of samples represent ethnic outliers in the study population, using Eigenstrat with iterative outlier removal, and reconciling these individuals with other ancestry information such as self-report could be used to identify a coherent set of ethnic outliers (which are identified as outliers *and* self-report as outliers) to potentially exclude from the analysis, rather than adjusting for many eigenvectors during the analysis only to retain a very small number of samples.

Sample genotyping efficiency / call rate

Genotyping efficiency, or call rate, is an issue which will be discussed in greater depth in the Marker Quality section below. A large proportion of SNP assays failing on an individual DNA sample may be indicative of a poor quality DNA sample, which could lead to aberrant genotype calling. Samples with low genotyping efficiency, or call rate, should be eliminated from further analysis. A recommended threshold is 98-99% efficiency, after first removing markers which have a low genotype call rate across samples. The suggested 98-99% threshold is an approximate threshold – the exact threshold may vary from study to study depending on the genotyping platform used, quality of the DNA samples used, and the variability in human and equipment error in genotyping. The threshold should be determined based on a goal whereby a balance minimizing the number of samples dropped and maximizing genotyping efficiency is attained. Figure 6 shows the proportion of samples (red and blue lines) or SNPs (green line) remaining at different call rate thresholds. Genotyping efficiency can be checked using the --missing option in PLINK. This will produce a file showing genotype missingness rate (1-efficiency) for each individual (proportion of SNPs which failed on each sample), and for each SNP (proportion of individuals for which no genotype was called). Samples below a desired threshold can be eliminated from any downstream analyses by using the --mind option in PLINK. Genotyping efficiency is also an important marker QC step, and is discussed below.

MARKER QUALITY

Marker genotyping efficiency / call rate

As mentioned in the sample genotyping efficiency section above, marker genotyping efficiency (the proportion of samples with a genotype call for each marker) is a good indicator of marker quality. SNP assays that failed on a large number of samples are poor assays, and are likely to result in spurious data. A recommended threshold for removing SNPs with low call rate is approximately 98-99%, although as mentioned in the sample genotyping efficiency section, this threshold may vary from study to study. Marker genotyping efficiency can be reviewed using the --missing option in PLINK. We recommend removing poor quality SNPs before running the sample genotyping efficiency check discussed above, so that fewer samples will be dropped from the analysis simply because they were genotyped with SNP assays that had poor performance (see Figure 6). Markers can be removed based on call rate by using the --geno option, followed by a threshold for a lower limit of missingness (e.g., --geno 0.02 would remove SNPs with more than 2% missing, i.e. less than a 98% call rate).

Control sample reproducibility / HapMap concordance

It is advantageous to incorporate internal controls in the genotyping pipeline to estimate genotyping reproducibility rate and for selecting which markers to eliminate based on poor reproducibility. Many studies routinely genotype DNA samples from the HapMap cell lines [39,40]. In addition to providing samples of known ancestry to anchor the Structure analysis discussed in the *Population Substructure* section above, genotype calls on HapMap samples can be compared to the corresponding publicly available reference genotypes to estimate the degree of concordance. Genotyping for the Marshfield PMRP and Group Health was performed by CIDR, which considered any SNP having more than one replicate error on HapMap samples run with the study samples to be a technical failure, and only intensity data were released for these markers. CIDR also considered SNPs technical failures if the SNP had a call rate <85%, if the absolute difference in call rate between sexes is greater than 2.5%, if the absolute difference in heterozygosity between sexes is greater than 7%, or if cluster separation <0.20. Vanderbilt BioVU, Mayo, and Northwestern NUGene samples were genotyped at the Broad Institute, where technical failure was determined by call rate

95%, GenTrain score <0.6 (a statistical measure from Illumina's clustering algorithm[41]), cluster separation <0.4, or more than one replicate error. It is also advantageous to build in duplicate samples to estimate the reproducibility rate within genotyping batches. By design, both CIDR and Broad include HapMap control samples and duplicate samples across all plates in the study. It is anticipated that for accurate genotyping data, duplicate reproducibility and HapMap concordance of >99% are expected. We removed any SNPs which had one or more discordant calls on duplicate samples. Both HapMap concordance and replicate sample concordance can be checked using the concordance procedure in the PLATO software [42] (Table 1) or by using the --genome --rel-check options in PLINK. Using HapMap trio samples it is also possible to inspect each SNP for Mendelian inconsistencies, which indicate genotyping errors if pedigree information is correct. Mendelian inconsistencies can be assessed using the --mendel option in PLINK. PLINK only detects Mendelian inconsistencies in full trios. The Mendelian-error procedure in PLATO will also evaluate Mendelian consistency in parent-offspring pairs or sib-pairs with missing parental genotypes.

We recommend removing or flagging any SNPs that have one or more Mendelian errors on HapMap control samples. While it may be possible to look for Mendelian inconsistencies using study samples, removing these SNPs could potentially be filtering out a phenotype specific copy number variant. If this is the case, there will likely be more than three genotype clusters. The extra clusters or parts of them will be missing or miscalled. For instance, for a locus with alleles A and B, A-, AA, and AAB may all cluster together unless the SNP is re-called with a specific model in mind.

Minor allele frequency

It is also important to filter SNPs based on minor allele frequency because statistical power is extremely low for rare SNPs. Figure 7 shows that the power to detect an association in a large dataset (n=10,000) with a relatively large effects (OR between 1.3 and 1.7) is extremely low for rare SNPs (<1% frequency).We recommend removing any extremely rare SNPs (including any monomorphic SNPs). The threshold chosen depends on the size of the study and the effect sizes expected. Power calculation software such as CaTS Power [43] or Quanto [44]can simplify power calculations for genetic association studies and inform the investigator of the allele frequency below which the study becomes severely underpowered. Minor allele frequency can be reported for each SNP using the --freq option in PLINK, and SNPs can be removed from the analysis using the --maf option, followed by a lower limit threshold. SNPs with frequency too low to yield reasonable statistical power (e.g. below 1%) may be removed from the analysis to lighten the computational and multiple testing correction burden. However, in studies with very large sample sizes it may be beneficial to avoid removing these rare SNPs. Others have shown that nonsynonymous, possibly deleterious SNPs are on average rarer than synonymous SNPs that likely do not cause any adverse phenotypes [45].

Hardy-Weinberg Equilibrium

Checking for Hardy-Weinberg Equilibrium (HWE) is one final step in the quality control analysis of markers in GWAS data. Under Hardy-Weinberg assumptions, allele and genotype frequencies can be estimated from one generation to the next. Departure from this equilibrium can be indicative of potential genotyping errors, population stratification, or even actual association to the trait under study [46]. UNIT 1.18 contains a very detailed description of the key principles and assumptions of HWE and how HWE is tested and applied in genetic association studies [47]. HWE can be assessed using the --hardy option in PLINK. While departure from HWE can indicate potential genotyping error, disequilibrium can also result from a true association. It has been consistently noted that many more SNPs

are out of HWE at any given significance threshold than would be expected by chance. SNPs severely out of HWE should therefore not be eliminated from the analysis, but flagged for further analysis after the association analyses are performed. Databases such as MySQL (see Table 1) can be very useful for joining association statistics with HWE statistics for easy reporting. It is also beneficial to examine HWE in controls separately, as disease-free controls should more closely follow the assumptions that lead to HWE than cases, and because some true associations are expected to be out of HWE. If multiple ethnicities are used in the same study it is necessary to test for HWE within each group separately. SNPs that are highly associated with the trait of interest that also show highly significant departures from HWE, especially in controls, should be closely scrutinized. Typically HWE deviations toward an excess of heterozygotes reflect a technical problem in the assay, such as non-specific amplification of the target region. On many GWAS platforms the quantitative allelic signals at a marker, i.e., the intensity plot for the SNP, can be used to screen for a technical origin of the HWE deviation: null alleles can produce multimodal genotype clusters in the heterozygote clusters and one of the homozygote clusters (Figure 8) or can produce an unexpected number of samples with no signal (Figure 9), and SNPs within CNVs or segmental duplications can produce clusters of genotypes intermediate between the three expected clusters of genotype (Figure 10) [48]. In the loci depicted in figures 8, 9 and 10, chi-square tests for HWE are rejected at P-values less than 10^{-80}, so these represent the most egregious examples of the aforementioned behavior. If no technical errors are detected then a number of biologically plausible explanations exist for HWE deviations toward an excess of homozygotes: population stratification, assortative mating and inbreeding, to name a few.

BATCH EFFECTS

Thousands of DNA samples are typically genotyped in a GWAS, which necessitates partitioning samples into small batches of samples processed in the lab together for genotyping (e.g. the set of samples on a 96 well plate). The precise size and composition of the sample batch depend on the array and lab process used. Systematic differences among the composition of individuals in a batch (i.e. the case to control ratio or race/ethnicity of individuals on plates) and the within-plate accuracy and efficiency can result in batch effects – apparent associations confounded by batch. The problem is in essence the same problem observed with population stratification – namely that if there is an imbalance of cases and controls on a plate, and there are nonrandom (unknown) biases or inaccuracies in genotyping that differ from plate to plate – spurious associations will result.

Ideally, no batch effect will be present because individuals with different phenotypes, sex, race, and other confounders should be plated randomly, and because modern high-throughput genotyping technology is much more accurate, efficient, and consistent than earlier generations of GWAS assays. There are several approaches for examining a dataset for potential batch effects. One simple approach is to calculate the average minor allele frequency and average genotyping call rate across all SNPs for each plate. Gross differences in either of these on any plates can easily be identified. Another method involves coding case/control status by plate followed by running the GWAS analysis testing each plate against all other plates. For example, the status of all samples on plate or batch 1 will be coded as case, while the status of every other sample is to be coded control. A GWAS analysis is to be performed (e.g. using the --assoc option in PLINK), and both the average p-value and the number of results significant at a certain threshold (e.g. $p<1\times10^{-4}$) can be recorded. SNPs with low minor allele frequency (i.e. <5%) should be removed before this analysis is performed to improve the stability of test statistics. This procedure should be repeated for each plate or batch in the study. If any single plate has many more or many fewer significant results, or has an average p-value that deviates from 0.5 (under the null the

average p-value will be 0.5 over many tests), then this batch should be further investigated for genotyping or composition problems. If batch effects are present, methods similar to those used for population stratification (e.g. genomic control) may be used to mitigate the confounding effects.

EVALUATION OF QC AFTER ASSOCIATION ANALYSIS

After phenotypic association analysis, the quality control measures used should be evaluated. One method is to compare the observed number of statistically significant results with a phenotype to expectation. Too many significant results may indicate insufficient QC. Also because no QC will catch all problematic SNPs, the intensity plots for statistically significant SNPs must be reviewed to make sure there are no obvious clustering problems. Replication of results using different genotyping technology (such as TaqMan) and/or in another sample may be needed as well.

FUTURE DIRECTIONS

The QC pipeline developed by the eMERGE network has enabled a thorough analysis of the quality of the genome-wide genotype data generated on the ~17,000 samples. All of these data have been deposited in dbGaP along with corresponding quality control documents that describe all of the QC details for each dataset individually. Conducting QC in parallel at a coordinating center and study sites has been a tremendously valuable experience, as it led to a more thorough understanding since each group had to reconcile its results with others.

Acknowledgments

This research was supported in part by NIH grants U01HG004608, U01HG004609, U01HG004610, U01HG04599, U01HG04603, U01HG004438, R01LM010040, and by the Intramural Research Program of the NIH, National Library of Medicine.

REFERENCES

1. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, Sangiovanni JP, Mane SM, Mayne ST, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005; 308:385–389. [PubMed: 15761122]
2. Frayling TM. Genome-wide association studies provide new insights into type 2 diabetes aetiology. Nat Rev Genet. 2007; 8:657–662. [PubMed: 17703236]
3. Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, Coin L, Najjar SS, Zhao JH, Heath SC, Eyheramendy S, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet. 2009
4. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, Kaplan L, Bennett D, Li Y, Tanaka T, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat Genet. 2009; 41:56–65. [PubMed: 19060906]
5. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, Clarke R, Heath SC, Timpson NJ, Najjar SS, Stringham HM, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet. 2008; 40:161–169. [PubMed: 18193043]
6. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009; 106:9362–9367. [PubMed: 19474294]
7. Barber MJ, Mangravite LM, Hyde CL, Chasman DI, Smith JD, McCarty CA, Li X, Wilke RA, Rieder MJ, Williams PT, et al. Genome-wide association of lipid-lowering response to statins in combined study populations. PLoS One. 2010; 5:e9763. [PubMed: 20339536]
8. Daly AK, Donaldson PT, Bhatnagar P, Shen Y, Pe'er I, Floratos A, Daly MJ, Goldstein DB, John S, Nelson MR, et al. HLA-B*5701 genotype is a major determinant of drug-induced liver injury due to flucloxacillin. Nat Genet. 2009; 41:816–819. [PubMed: 19483685]

9. Link E, Parish S, Armitage J, Bowman L, Heath S, Matsuda F, Gut I, Lathrop M, Collins R. SLCO1B1 variants and statin-induced myopathy--a genomewide study. N Engl J Med. 2008; 359:789–799. [PubMed: 18650507]
10. Thompson JF, Hyde CL, Wood LS, Paciga SA, Hinds DA, Cox DR, Hovingh GK, Kastelein JJ. Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort. Circ Cardiovasc Genet. 2009; 2:173–181. [PubMed: 20031582]
11. 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; 39:1181–1186. [PubMed: 17898773]
12. McCarty C, Chrisolm R, Chute C, Kullo I, Jarvik G, Larson E, Li R, Masys D, Ritchie M, Roden D, et al. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. BMC Medical Genomics. 2010 In Revision.
13. Laurie C, Mirel D, Pugh E, Bierut L, Bhangale T, Boehm F, Caporaso N, Edenburgh H, Gabriel S, Harris E, et al. Quality control and quality assurance in genotypic data for genome-wide association studies. Genetic Epidemiology. 2010 In Press.
14. Miyagawa T, Nishida N, Ohashi J, Kimura R, Fujimoto A, Kawashima M, Koike A, Sasaki T, Tanii H, Otowa T, et al. Appropriate data cleaning methods for genome-wide association study. J Hum Genet. 2008; 53:886–893. [PubMed: 18695938]
15. Chanock SJ, Manolio T, Boehnke M, Boerwinkle E, Hunter DJ, Thomas G, Hirschhorn JN, Abecasis G, Altshuler D, Bailey-Wilson JE, et al. Replicating genotype-phenotype associations. Nature. 2007; 447:655–660. [PubMed: 17554299]
16. Broman KW. Cleaning genotype data. Genet Epidemiol. 1999; 17(Suppl 1):S79–S83. [PubMed: 10597416]
17. 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; 81:559–575. [PubMed: 17701901]
18. Illumina Technical Note: “TOP/BOT” Strand and “A/B” Allele. 2009. http://pngu.mgh.harvard.edu/~purcell/plink/
19. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–311. [PubMed: 11125122]
20. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2005. ISBN 3900051070, URL http://www.R-project.org
21. Turner, SD. Visualizing sample relatedness in a GWAS using PLINK and R. 2009. https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Visualizing_relatedne ss
22. Aulchenko YS, de Koning DJ, Haley C. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. Genetics. 2007; 177:577–585. [PubMed: 17660554]
23. Cardon LR, Palmer LJ. Population stratification and spurious allelic association. Lancet. 2003; 361:598–604. [PubMed: 12598158]
24. Census 2000: Profile of Demographic Characteristics, Marshfield WI. 2000. http://censtats.census.gov/data/WI/1605549675.pdf
25. Tang H, Quertermous T, Rodriguez B, Kardia SL, Zhu X, Brown A, Pankow JS, Province MA, Hunt SC, Boerwinkle E, et al. Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. Am J Hum Genet. 2005; 76:268–275. [PubMed: 15625622]
26. Dumitrescu LC, Ritchie MD, Brown-Gentry K, Pulley JJ, Basford M, Denny J, Oksenberg JR, Roden DM, Haines JL, Crawford DC. Assessing the accuracy of observer-reported ancestry in a biorepository linked to electronic medical records. Genetics in Medicine. 2010 In Press.
27. Manolio TA. Collaborative genome-wide association studies of diverse diseases: programs of the NHGRI’s office of population genomics. Pharmacogenomics. 2009; 10:235–241. [PubMed: 19207024]
28. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999; 55:997–1004. [PubMed: 11315092]

29. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol. 2001; 20:4–16. [PubMed: 11119293]
30. Dadd T, Weale ME, Lewis CM. A critical evaluation of genomic control methods for genetic association studies. Genet Epidemiol. 2009; 33:290–298. [PubMed: 19051284]
31. Devlin B, Bacanu SA, Roeder K. Genomic Control to the extreme. Nat Genet. 2004; 36:1129–1130. [PubMed: 15514657]
32. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000; 155:945–959. [PubMed: 10835412]
33. STRUCTURE. 2009. http://pritch.bsd.uchicago.edu/structure.html
34. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004; 36:512–517. [PubMed: 15052271]
35. 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; 38:904–909. [PubMed: 16862161]
36. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006; 2:e190. [PubMed: 17194218]
37. Zhang F, Wang Y, Deng HW. Comparison of population-based association study methods correcting for population stratification. PLoS One. 2008; 3:e3392. [PubMed: 18852890]
38. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, et al. Genes mirror geography within Europe. Nature. 2008; 456:98–101. [PubMed: 18758442]
39. International hapmap consortium. The International HapMap Project. Nature. 2003; 426:789–796. [PubMed: 14685227]
40. International hapmap consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–861. [PubMed: 17943122]
41. Illumina GenCall Data Analysis Software. 2008. http://www.illumina.com/Documents/products/technotes/technote_gencall_data_analysis_software.pdf
42. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, Ritchie MD. Finding unique filter sets in plato: a precursor to efficient interaction analysis in gwas data. Pac Symp Biocomput. 2010:315–326. [PubMed: 19908384]
43. 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; 38:209–213. [PubMed: 16415888]
44. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med. 2002; 21:35–50. [PubMed: 11782049]
45. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet. 2008; 82:100–112. [PubMed: 18179889]
46. Wittke-Thompson JK, Pluzhnikov A, Cox NJ. Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet. 2005; 76:967–986. [PubMed: 15834813]
47. Ryckman K, Williams SM. Calculation and use of the Hardy-Weinberg model in association studies. Curr Protoc Hum Genet. 2008 Chapter 1: Unit.
48. Carlson CS, Smith JD, Stanaway IB, Rieder MJ, Nickerson DA. Direct detection of null alleles in SNP genotyping data. Hum Mol Genet. 2006; 15:1931–1937. [PubMed: 16644863]

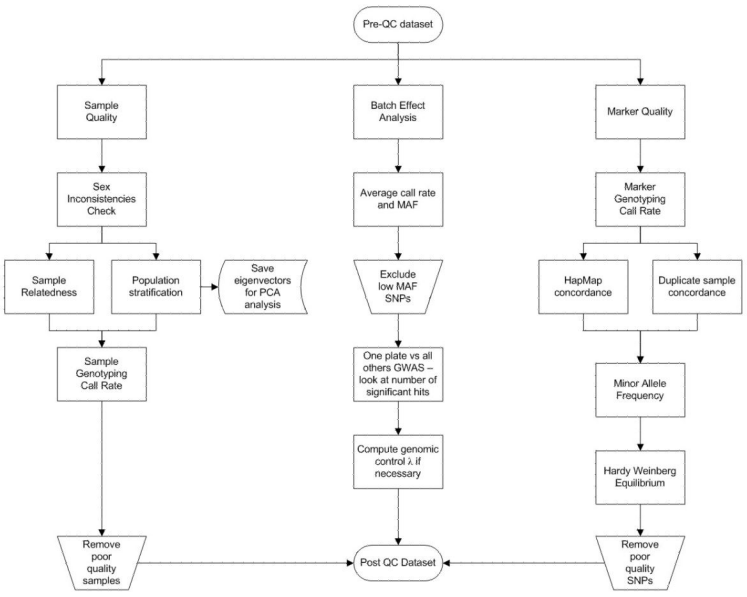

Figure 1.
A flowchart overview of the entire GWAS QC process. Each topic is discussed in detail in the corresponding section in the text. Squares represent steps, ovals represent input or output data, and trapezoids represent filtering of data.

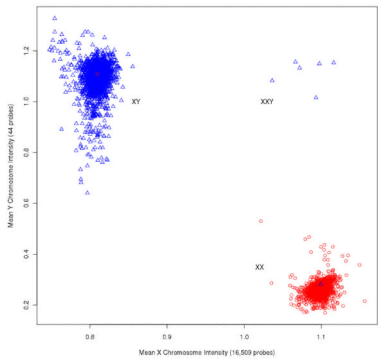

Figure 2.

Visualization of X and Y probe intensities. The x-axis and y-axis represent the sum of the average over all probes for the normalized Cartesian intensity for allele A and the average over all probes for the normalized Cartesian intensity for allele B using all probes available on X chromosome and Y chromosome, respectively. The XX (female, red circles) and XY (male, blue triangles) subjects are shown on the bottom right corner and on the top left corner, respectively. The plot reveals two mislabeled individuals (one male with the female cluster, and one female with the male cluster). Several XXY individuals are also clearly visible (upper right corner).

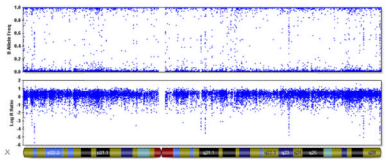

Figure 3.

Copy Number and allelic variation to detect anomalies on X chromosome. The top plot shows the B-Allele frequencies for all probes for one sample with total loss of heterozygosity (LOH) on X chromosome. The bottom plot shows the copy number variation from the same sample on X chromosome. Both plots are helpful to detect regions of LOH and/or copy number variation such as deletion and amplification.

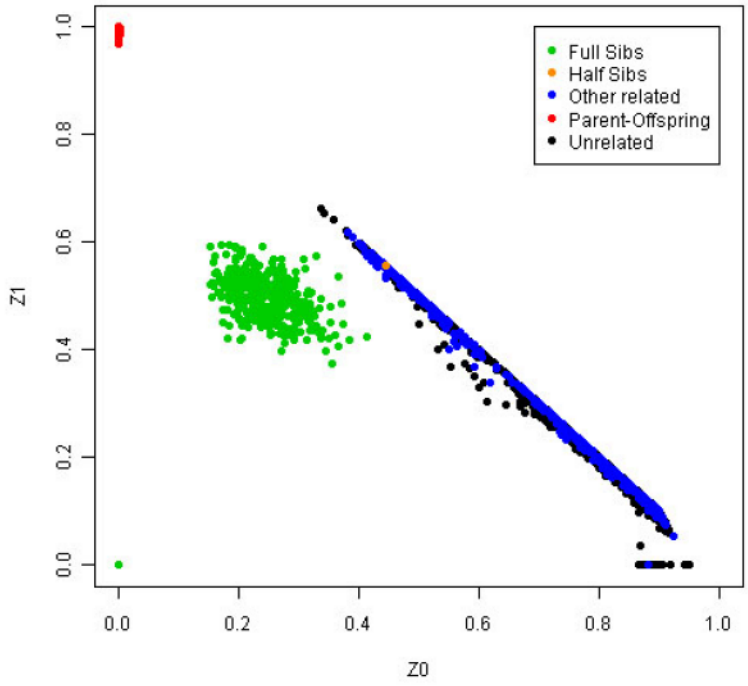

Figure 4.

Points in this plot show pairs of individuals plotted by their degree of relatedness: the proportion of loci where the pair shares one allele IBD (Z1) by the proportion of loci where the pair shares zero alleles IBD (Z0). These values are obtained from PLINK using the --genome option. Pairs are color-coded by the type of relationship determined by the pedigree information embedded in the pedfile (also reported by PLINK). This plot omits pairs of individuals having an overall kinship coefficient ≥ 0.05 for clarity. There is a pair of monozygotic twins represented by a point in the lower left at (0,0), because they share two alleles IBD at every locus across the genome.

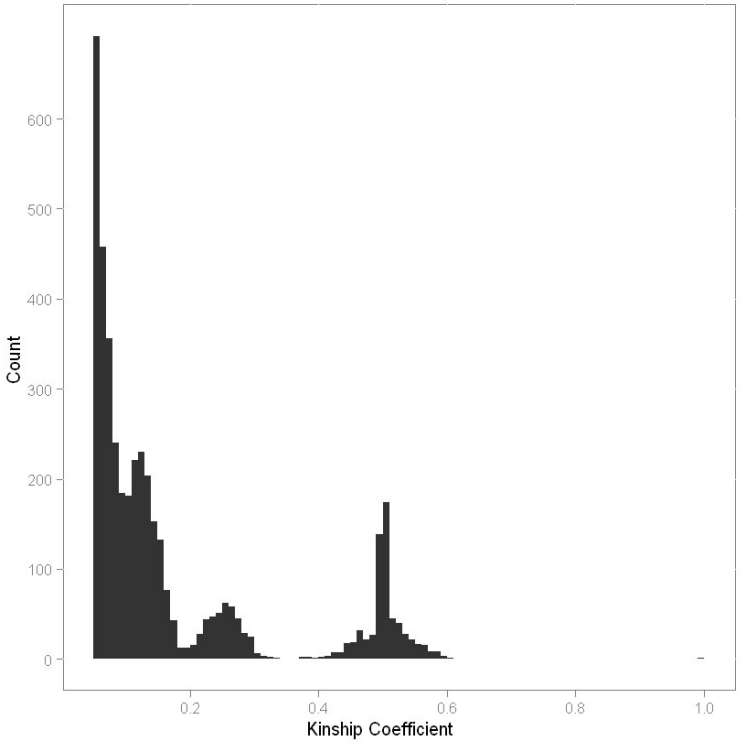

Figure 5.
Histogram showing the distribution of pairwise kinship coefficients (where kinship coefficient is greater than 0.05). The peak over 0.5 represents first degree relatives (parent-offspring, full siblings). The peak over 0.25 represents second degree relatives (half siblings, avuncular, grandparent-grandchild). Third and fourth degree relatives begin to blend into more distantly related samples between zero and 0.125.

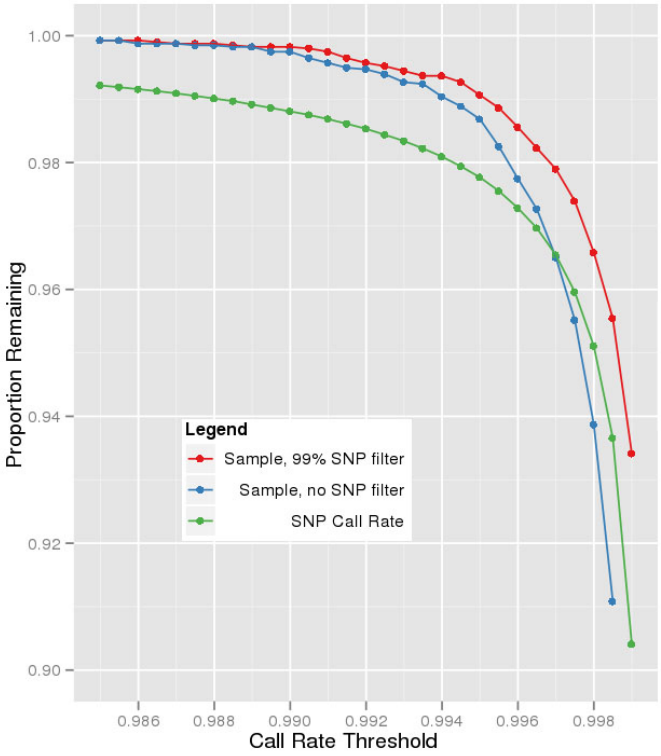

Figure 6.
Proportion of SNPs or samples remaining as call rate threshold increases. The green line shows the propotion of SNPs remaining when SNPs are discarded if they fall below the given genotyping efficiency threshold. The blue line shows the proportion of samples remaining, while the red line shows the proportion of samples remaining if a 99% call rate threshold is applied to eliminate poor quality markers first.

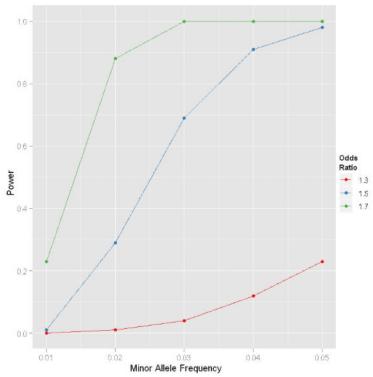

Figure 7.

This shows the power to detect an association at genome-wide significance ($p<5\times10^{-8}$), assuming the actual causal SNP is genotyped in a case-control study consisting of 5000 cases and 5000 controls of a common disease with 10% prevalence under an additive model at several different odds ratios. Note that when the MAF is low, power is extremely low even for very large effects (OR=1.7).

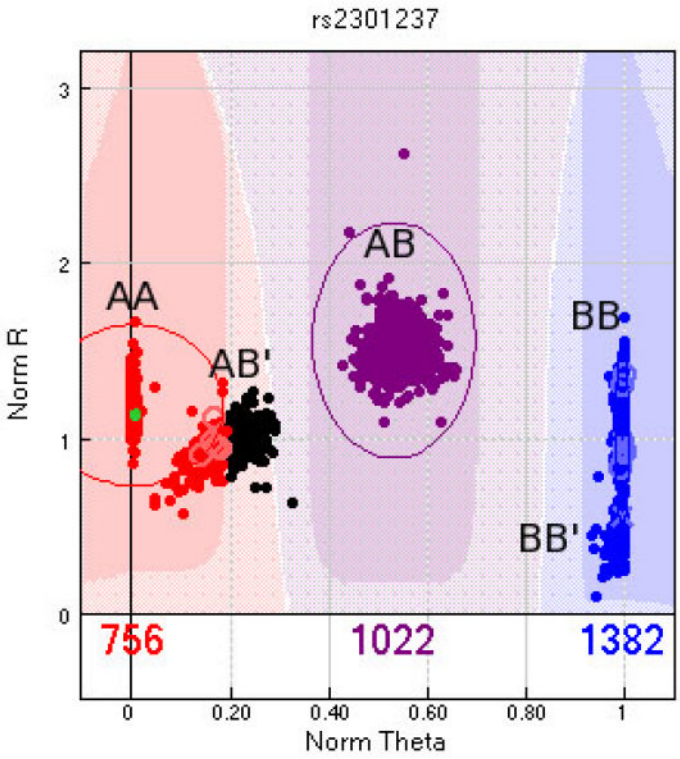

Figure 8.
AB and BB individuals are split into sub-clusters AB and AB', BB and BB', while AA cluster is unaffected. The AB/AB' split results in some AB samples miscalled as AA (diagnosed by Mendelian inconsistencies in the genotypes), as well deviation from HWE due to excess homozygosity. Since only samples with at least one B allele demonstrate the splitting, one consistent explanation is the presence of a cryptic polymorphism near rs2301237 on a haplotype that contains the B allele. In this case, a second polymorphism (rs3114267) lies eight bases upstream from the typed polymorphism, and is in complete LD (D'=1, r^2=.2) with rs2301237.

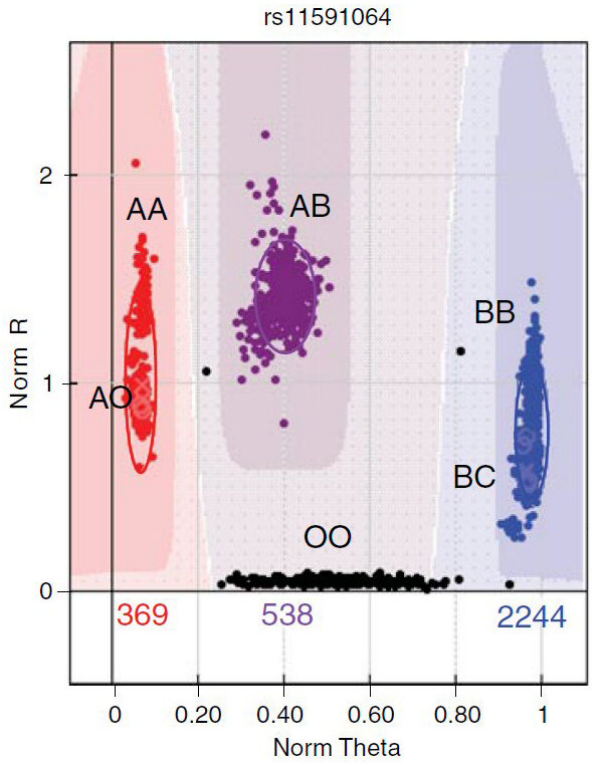

Figure 9.

Unexpected number of clusters resulting in departure from HWE consistent with copy loss. Hemizygous individuals cluster at AO and BO. Individuals with homozygous deletions cluster at OO and their genotype calls are missing. The AB cluster remains intact, since these individuals are *ipso facto* diploid at the locus. Parent-parent-child Mendelian errors are present when at least one parent is hemizygous and produces hemizygous offspring. The deletion results in excess homozygosity. In this case, the “copy loss” appears to be a six-nucleotide insertion (rs71578153) coincident with rs11591064 that disrupts both A and B probes.

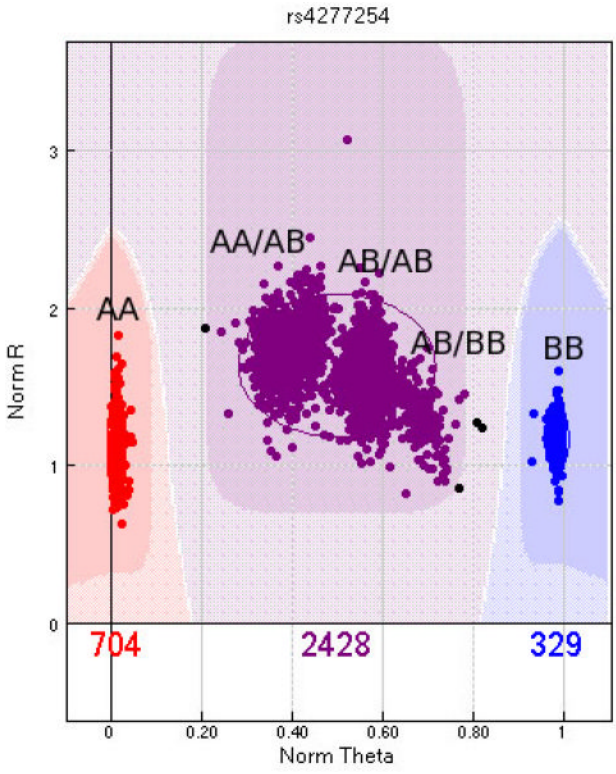

Figure 10.
The five observed clusters are most consistent with a segmental duplication, although none is curated around the locus. A copy number variant would be expected to produce additional clusters above the AA and BB clusters (ie, AAA and BBB), as opposed to the splits being confined to strictly the heterozygous clusters. Regardless, the artifact results in excess heterozygosity.

Table 1

Useful software packages for data management, quality control, and statistical analysis in genome-wide association studies.

Software package	URL	Purpose
PLINK	http://pngu.mgh.harvard.edu/~purcell/plink/	Free, open-source GWAS analysis software package. Contains many tools for data management, quality control, and statistical analysis. (PC, Mac, Linux).
PLATO	https://chgr.mc.vanderbilt.edu/plato	*PL*atform for the *A*nalysis, *T*ranslation, and *O*rganization of large-scale data – software for GWAS analysis similar to PLINK.
R	http://www.r-project.org/	Free, open-source statistical computing software with excellent graphical capabilities. (PC, Mac, Linux).
Eigensoft	http://genepath.med.harvard.edu/~reich/Software.htm	Free, open-source software for performing principal components analysis based method for detecting and correcting for population stratification in GWAS. (Linux only).
Structure	http://pritch.bsd.uchicago.edu/structure.html	Free, open-source software for inferring the presence of distinct populations and assigning individuals to those populations for a stratified analysis. (Windows, Mac, Linux)
MySQL Workbench	http://wb.mysql.com/	Free, open-source software for creating, administering and querying relational databases. This is helpful for subsetting data, merging results, and joining QC metrics (e.g. HWE) to final association results. (Windows, Mac, Linux).

Table 2

Example table showing output from --check-sex routine using PLINK. IID=individual id; PEDSEX=sex as recorded in pedfile (1=male, 2=female); SNPSEX=sex as predicted based on genetic data (1=male, 2=female, 0=unknown).

IID	PEDSEX	SNPSEX	STATUS	F	Explanation
1	1	1	OK	0.98	Male
2	2	2	OK	0.03	Female
3	2	1	PROBLEM	0.99	Recorded female, genetically male
4	1	2	PROBLEM	0.02	Recorded male, genetically female
5	2	0	PROBLEM	0.28	Likely a female with sex chromosome anomaly (e.g. XX/XO mosaic, loss-of-heterozygosity on X)
6	1	0	PROBLEM	0.35	Likely a male with sex chromosome anomaly (e.g. XXY or XX/XY mosaic)