Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure

Yik Y. Teo[a,b]

[a]Wellcome Trust Centre for Human Genetics, University of Oxford, UK and [b]Wellcome Trust Sanger Institute, Hinxton, UK

Correspondence to Yik Y. Teo, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK
Tel: +44 1865 287712; fax: +44 1865 287501;
e-mail: teo@well.ox.ac.uk

Current Opinion in Lipidology 2008, 19:133–143

Purpose of review

Genetic association studies which survey the entire genome have become a common design for uncovering the genetic basis of common diseases, including lipid-related traits. Such studies have identified several novel loci which influence blood lipids. The present review highlights the statistical challenges associated with such large-scale genetic studies and discusses the available methodological strategies for handling these issues.

Recent findings

The successful analysis of genome-wide data assayed on commercial genotyping arrays depends on careful exploration of the data. Unaccounted sample failures, genotyping errors and population structure can introduce misleading signals that mimic genuine association. Careful interpretation of useful summary statistics and graphical data displays can minimize the extent of false associations that need to be followed up in replication or fine-mapping experiments.

Summary

Recently published genome-wide studies are beginning to yield valuable insights into the importance of well designed methodological and statistical techniques for sensible interpretation of the plethora of genetic data generated.

Keywords

coverage, genome-wide association study, genotyping error, population structure, power, quality control

Curr Opin Lipidol 19:133–143

0957-9672

Introduction

Genome-wide association study (GWAS) is increasingly common as an experimental design for investigating the genetic basis of common diseases and complex traits in humans. Such study designs have been made possible by extensive databases on human genetic variations [1–3,4••] and advances in genotyping technologies. The development of sophisticated bioinformatics and statistical tools has also been vital to manage, analyze and interpret the plethora of genetic and epidemiological data [5–8,9•, 10,11••,12•,13••,14••,15•]. Faced with potentially a million single nucleotide polymorphisms (SNPs), in which the true signal of phenotypic association may not be substantially larger than background noise or confounding effects, the statistical challenges presented in GWAS can be significantly different from conventional clinical trials [16,17]. While traditional issues related to the design and conduct of an experiment still exist (for example, sample size calculation, multiple testing, and confounding), the extent of these problems is compounded in GWAS given the number of variables investigated and large sample sizes. This review will discuss the practical issues related to the design and analysis of data for GWAS, and provide a brief overview of the range of statistical tools available for addressing these concerns.

Experimental design

Almost all the recent publications in GWAS have used a case–control design, which compares a set of unrelated individuals with the trait of interest against an unaffected and unrelated set of controls [18–26]. The relative ease of recruiting a large number of participants for a case–control study makes it an attractive alternative to family-based designs, in which adequate recruitment can often be difficult, especially in the context of late-onset diseases [27]. Sample dropouts due to genotyping failures, relationship misspecifications and laboratory errors can be more costly in family-based studies, since the loss of one individual may potentially result in the exclusion of an entire pedigree. McGinnis and colleagues [28] have shown in earlier work that comparable power is obtained for the same number of case–control pairs and nuclear trios. This

means a study design using trios will cost 50% more than a case–control design for similar statistical power, given the additional genotyping that has to be performed.

A main disadvantage of case–control study designs is the vulnerability to confounding caused by undetected or unaccounted population structure [29–34], which family-based studies are less susceptible to. Sophisticated statistical techniques, however, have been successfully developed to detect and correct for the effects of population structure, which will be discussed in greater extent at a later stage of this review.

One argument in favor of using family-based studies is the greater ability to detect rare variants compared with case–control studies. As a family pedigree is recruited when at least one member of the family is ascertained to possess the trait of interest, there will be a greater concentration of disease-predisposing alleles in the pedigree, relative to the general population. This means there is a higher likelihood of observing an association between the transmission of a disease predisposing allele and trait onset. The environmental exposure for members within a pedigree is also more homogeneous, and this minimizes nongenetic differences in disease architecture which are attributed to diet, lifestyle and the environment. Conversely, it can be difficult to separate nongenetic and genetic causes when using unrelated individuals due to the heterogeneous environmental exposure between the participants.

Power and coverage

The ability to identify genetic variants that genuinely result in phenotypic variations is defined as the power of the experiment. Optimizing the power without inflating the rate of false positive discovery (defined as an erroneous association with the trait of interest) is important when designing a GWAS, since the number of genetic markers assayed in a typical GWAS can be over 100 000 and a large number of putative associations may occur simply by chance if conventional definition of statistical significance is used. Even at a seemingly stringent significance threshold of 0.001, we expect 500 SNPs to be associated due to chance alone (thus without any biological relevance) when genotyping on a platform which assays 500 000 markers. Bonferroni correction is a common procedure for guarding against the increased likelihood of obtaining a significant result due to chance, and is implemented by dividing the nominal significance threshold by the number of independent tests performed to yield a more stringent criterion for assessing each hypothesis test. This procedure can be overly conservative in GWAS since the SNPs tested are correlated (cf. linkage disequilibrium later) and thus the total number of independent tests is often less than the number of assayed SNPs [35]. As the use of overly stringent significance thresholds requires stronger statistical evidence before a trait-associated SNP is considered statistically meaningful (either in terms of a larger effect size, or through the use of larger sample sizes), the choice of the significance threshold can thus affect the power of the study. Most recent studies have adopted statistical significances between 10^{-4} and 10^{-7} to minimize false associations [13••, 18–26], although the interpretation of SNPs with evidence stronger than these criteria ranges from 'moderate' to 'strong' evidence for association [13••].

The coverage of a genotyping platform refers to the extent of genetic variation in the human genome that has been represented by the markers on the platform. An untyped SNP which is strongly correlated (conventionally defined as a correlation, or linkage disequilibrium, of more than 0.8) with the markers on the platform is therefore defined to be 'tagged'. Unless every polymorphic marker in the genome is documented, current discussion of coverage is inevitably restricted to evaluating whether the extent of common genomic variation, as identified by the International HapMap Project [3,4••], has been successfully characterized [36,37••,38•]. An important note, however, is that different criteria for assessing coverage can yield dramatically different estimates. For the same platform, genomic coverage assessed using pairwise tagging strategies [39] will always be lower than multiloci or haplotype tagging [36,37••], since pairwise tagging only quantifies the correlation between a focal SNP and one tag SNP, while multiloci tagging considers the additional correlation between the alleles at a focal SNP and the haplotypes from surrounding multiple tag SNPs. For example, the Affymetrix 500K array has a genomic coverage of 67% and 80% when assessed using pairwise and three-marker tagging, respectively.

Recent publications have celebrated the success of genome-wide strategies for categorizing genetic variants which present unequivocal evidence for trait affiliation. These studies have mainly been restricted to populations of European descent. As there is substantially lesser linkage disequilibrium in African populations than in non-African populations [3,37••,40•], the coverage provided by genotyping platforms can vary dramatically when moving between non-African and African populations [37••,38•] (e.g. from 60.6% in the HapMap CEU to 37.2% in the HapMap YRI for the Affymetrix 500K chip).

The relation between coverage and power is commonly misunderstood in genetics, and researchers often mistake a lack of coverage to be the same as having low power. GWAS essentially relies on indirect association to identify regions containing the functional polymorphisms by testing markers which are in linkage disequilibrium with the functional polymorphisms rather than testing the functional polymorphisms directly. This means the

ability to detect a true biological association depends on the extent of linkage disequilibrium between the genotyped markers and the variants with the true effect. Current measures of pairwise or multiloci SNP relationships tend to be conservative, and markers not tagged or represented by conventional definitions may have complex affiliations with surrounding variants from which sophisticated methodologies can extract information [41–44]. This means that a SNP not classified as being tagged (by conventional definition of pairwise or multiloci r^2 greater than some threshold) may exist on particular haplotypic backgrounds that allow for accurate inference of the alleles at the SNP. Sufficient power thus can exist for detecting a variant which is not in high linkage disequilibrium with surrounding markers when assessed using conventional low-dimensional correlation measures. This was demonstrated in various applications combining clever use of statistics and available catalogues of common genetic variants [4••,13••,14••,45•,46], when associations in untyped regions with low reported coverage were detected using SNPs on available genotyping platforms. These techniques impute the genotypes for the untyped SNPs and provide a tool for in-silico fine mapping, thus increasing the power to detect and identify regions containing the causal variant.

Technological advancements can boost statistical power in a GWAS through careful design of dense genotyping arrays that prioritizes the selection of highly informative tag SNPs to improve coverage, while increasing the sample size of the study is another experimental procedure for augmenting power [47,48]. The genetic architecture for common diseases and complex traits is expected to depend on multiple genetic variants with small effects [16,17]. This necessitates the use of large sample sizes to elucidate the marginal influence each genetic variant has on the trait. While sample size calculations are typically performed to justify grant applications, the truth in this matter is there is a need to include as many samples as realistically possible given recruitment and financial constraints. A common strategy to increase power and authenticate putative associations is to perform in-silico replication, when data from several GWASs of the same trait are combined in a metaanalytic approach [19–21,49]. This pools the information from multiple studies, thus increasing the effective sample size. Variations in the design of the different studies, however, such as different phenotypic definitions, population structure and environmental exposure, can introduce systematic differences throughout the genome-wide comparisons and these need to be explicitly modeled to achieve meaningful results in the metaanalyses.

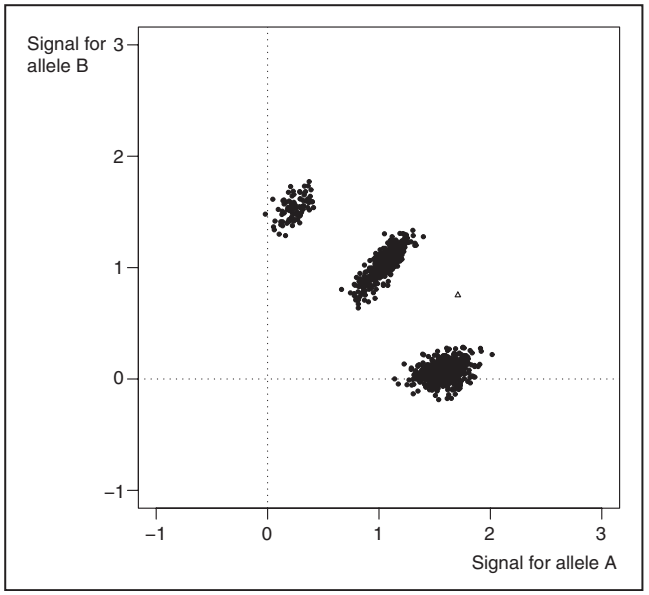
In a strict sense, calculations of power and coverage should always be made with respect to the underlying linkage disequilibrium present in the study population and the genotyping technology used. Dense genotyping arrays assaying well chosen markers can increase the number of SNPs with stronger correlations with underlying functional polymorphisms, thus increasing the statistical power in a GWAS. Power is also dependent on the allelic architecture of the disease, and there is lower power to identify rare compared with common variants. Sophisticated statistical methods for haplotypic analysis and directly imputing the functional polymorphisms will be useful to sharpen the weak association signals, but there is no better substitute than increasing the sample size when it comes to power considerations.

Genotype calling

Advances in genotyping technology have played a vital role in making GWAS realistically possible and affordable. Up to a million genetic variants can now be assayed simultaneously with predesigned oligonucleotide microarrays designed by either Affymetrix (Santa Clara, California, USA) or Illumina (San Diego, California, USA). These microarrays typically assay SNPs although the recent Affymetrix SNP Array 6.0 and Illumina Human 1M BeadChip are reported to profile copy number variants as well. DNA genotyping yields a series of hybridization intensities which need to be translated into the actual genotypes, through the process known as genotype calling.

Traditionally, genotypes were manually determined in the laboratory by examining fluorescent intensities of allelic expression. Automated procedures are necessary with the advent of large-scale genotyping, which assays at least hundreds of thousands of SNPs. These procedures are often unsupervised and typically rely on predefined rules to assign genotypes [13••,15•,50–55]. To account for inter-sample and batch variability in the extent of reagent washout and amount of input DNA, the raw hybridization intensities typically need to be normalized for meaningful comparisons between samples [13••,15•,52–57]. This is relevant since recently developed genotype calling algorithms typically adopt clustering strategies that combine information across different samples at each SNP and assign genotype calls to entire clusters simultaneously, compared with earlier strategies [51], which relied on the intensities for each individual sample at each SNP and are susceptible to preferentially assign heterozygous genotypes as missing. Pooling information across all the individuals to assign genotypes increases the confidence of the call, since samples with similar intensity profiles are expected to have the same genotype (see Fig. 1, which shows a visual representation of the allelic signals for a collection of individuals at a particular SNP, and such display is termed a clusterplot). This strategy, however, implies that the accuracy of the algorithm can depend on the number of samples within each genotype cluster, and calling genotypes for a small number of samples or for

Figure 1 A clusterplot for 1504 samples at a single nucleotide polmorphism (SNP)



The horizontal and vertical axes represent the normalized hybridization signals for the two possible alleles at a SNP, generically termed as alleles A and B, respectively. Each point corresponds to the signal profile for a sample and negative signals are possible for some normalization methods (for example, that adopted by the Wellcome Trust Case Control Consortium [13••]). The cluster of points which has a high signal for allele A but has a near-zero signal for allele B is expected to correspond to samples with the AA genotype. Conversely, the cluster of points with high signals for allele B but near-zero signals for allele A is expected to correspond to samples with the BB genotype. The cluster of points with almost similar signal profiles for alleles A and B is expected to correspond to samples with the heterozygous AB genotype. The sample represented by an open triangle has a signal profile that cannot be unambiguously assigned to one cluster, and well calibrated calling algorithms will assign a null or missing call to such an outlier. Most recently developed calling algorithms rely on cluster separation methods to assign genotype calls.

SNPs with rare alleles can potentially be more prone to errors (see section 'Quality control').

A number of independent software packages for calling genotypes have been developed for both Affymetrix and Illumina genotyping arrays. While the purpose of this article is not to review all the available genotype calling algorithms, we provide an overview of the software available and commonly used for the different genotyping technologies (Table 1). The technology and design for Affymetrix arrays have seen rapid developments, and these have necessitated modifications to the calling algorithms for the different platforms. By comparison, the technology from Illumina has remained fairly consistent, and thus there has been less modification to the calling algorithms.

These algorithms typically do not directly assign genotypes, but instead calculate metrics of confidence for each of the three possible genotype calls for a sample at every SNP. Recently developed calling algorithms (for example, Chiamo [13••], Illuminus [15•] and Xtyping [55]) estimate the probabilities for each of the three possible genotypes (AA, AB, BB) given the observed signal data across all the individuals. The user then defines a threshold for these probabilities. At a SNP, the most likely genotype is assigned to a sample if the corresponding probability is larger than the threshold, otherwise a null or missing genotype is assigned. Thus the user-defined threshold can be adjusted according to the desired compromise between accuracy and call rates. This means the extent of missing genotypes is dependent on the calling algorithm and the threshold used, and there is no reason to expect that the same probability threshold of 0.95 will result in similar performance across different calling algorithms. As these probabilities essentially quantify the degree of uncertainty in making a call, such information can be incorporated in downstream association analysis averaging over call uncertainty using missing data likelihood [14••].

Table 1 Current genotyping technologies and their respective calling algorithms

Platform	Programs available	Developers	Available online	Reference
Affymetrix				
GeneChip Human Mapping100 K Set	DM	Affymetrix	Yes[a]	[51]
	BRLMM	Affymetrix	Yes[a]	[53]
GeneChip Human Mapping500 K Set	DM	Affymetrix	Yes[a]	[51]
	BRLMM	Affymetrix	Yes[a]	[53]
	CHIAMO	Marchini *et al.*, WTCCC	Yes	[13••]
	XTYPING	Plagnol *et al.*	Yes	[55]
Genome-wide Human SNP Array 5.0	BRLMM-P	Affymetrix	Yes[b]	[54]
Genome-wide Human SNP Array 6.0	BIRDSEED	Affymetrix	Yes[b]	NA, see [54]
Illumina				
Sentrix HumanHap300 Genotyping BeadChip	GenTrain/GenCall	Illumina	Yes[c]	
Sentrix HumanHap550 Genotyping BeadChip	ILLUMINUS	Teo *et al.*	Yes	[15•]
Sentrix HumanHap650Y Genotyping BeadChip				
Human1M DNA Analysis BeadChip				

[a] Available on the Affymetrix GeneChip Operating Software (GCOS).
[b] Available on the Affymetrix Genotyping Console Software.
[c] Available on the Illumina BeadStudio analysis software package from Illumina Connect (http://www.illumina.com/IlluminaConnect).

The choice of the genotype calling algorithm is thus relevant to the type of downstream analyses that are possible or necessary.

In summary, the number of statistical approaches for managing and translating hybridization intensities into genotype calls is increasing with advancements in genotyping technology. As the field moves towards assaying millions of SNPs across tens of thousands of samples, computational speed and memory requirements of a calling algorithm become increasingly relevant in addition to accuracy and call rates. Ultimately, a genotype calling algorithm will only be useful if it is accurate with low rates of missing genotypes, fast and realistically usable by research groups with limited high-performance computing resources.

Quality control

Quality control refers to the exploratory procedures used to evaluate the genotyping performance of the samples and the genotyping array. As there can be degradation of input DNA, plating errors and hybridization failures of genotyping chips, it is important to review the performance of the samples prior to definitive downstream analysis with the genotypes. The process of calling genotypes is not error free, and differential performance can occur between SNPs. It is thus vital to identify and exclude SNPs with potentially high rates of missingness or erroneous genotypes. In this section, we provide a discussion on the quality control filters that are effective in minimizing the number of problematic samples and SNPs in subsequent analyses.

Sample quality control

The extent of missing genotypes and heterozygosity for a sample are useful indicators for poorly genotyped samples. Samples with anomalously high rates for either of these two measures are often excluded from the outset. High rates of missingness generally imply hybridization problems, which may be caused by faulty arrays or poor quality DNA; excess heterozygosity can indicate sample contamination, resulting in a disproportionate amount of heterozygous genotypes. It is often useful to represent missingness and heterozygosity in the same figure to decide on the filtering threshold (Fig. 2a). Unintentional use of related samples or accidental sample duplication can often occur in large-scale studies. Such cryptic relatedness is easy to infer through measures of allele sharing given the vast amount of genetic information, and typically the sample in each relation with the least amount of missing genotypes is retained in the study. For family-based studies, further assessment of the authenticity of the pedigree relationships can be achieved by calculating the extent of mendelian inconsistent genotypes across the genome, and samples that exhibit clear evidence of relationship misspecifications may have to be excluded or analyzed separately.

Single nucleotide polymorphism quality control

Quantile–quantile plots compare the obtained test statistics against what is expected under the null hypothesis of no association, and these are regularly used in SNP quality assessment. Quantile–quantile plots provide a tool for visualizing and assessing the extent of systematic deviation from the distribution under the null hypothesis (cf. genomic control later), and to identify outliers resulting from genotyping errors or potential association with the trait (see Fig. 2b). The extent of missing genotypes from well calibrated calling algorithms has been found to be a good surrogate for genotyping accuracy and SNP performance [13••]. Removing SNPs with a greater proportion of missing genotypes can yield a set of SNPs with accurate genotype calls for downstream analyses and improve the overall result of the study (Fig. 2b), since differential rates of missingness between cases and controls can produce spurious associations [58]. Genotype calling algorithms have the potential to make incorrect calls and such errors can be difficult to detect using the corresponding confidence metrics, especially when the entire SNP is poorly called (Fig. 2c). Checking for gross departure to Hardy–Weinberg equilibrium (HWE) has been shown to help in identifying SNPs with obvious genotyping errors [59]. As most clustering-based calling algorithms tend to perform poorly for SNPs with rare alleles, it is often useful to exclude SNPs with low minor allele frequencies (MAFs) from further analyses since they are generally less informative and current genome-wide designs are not powered to study such variants (Fig. 2b).

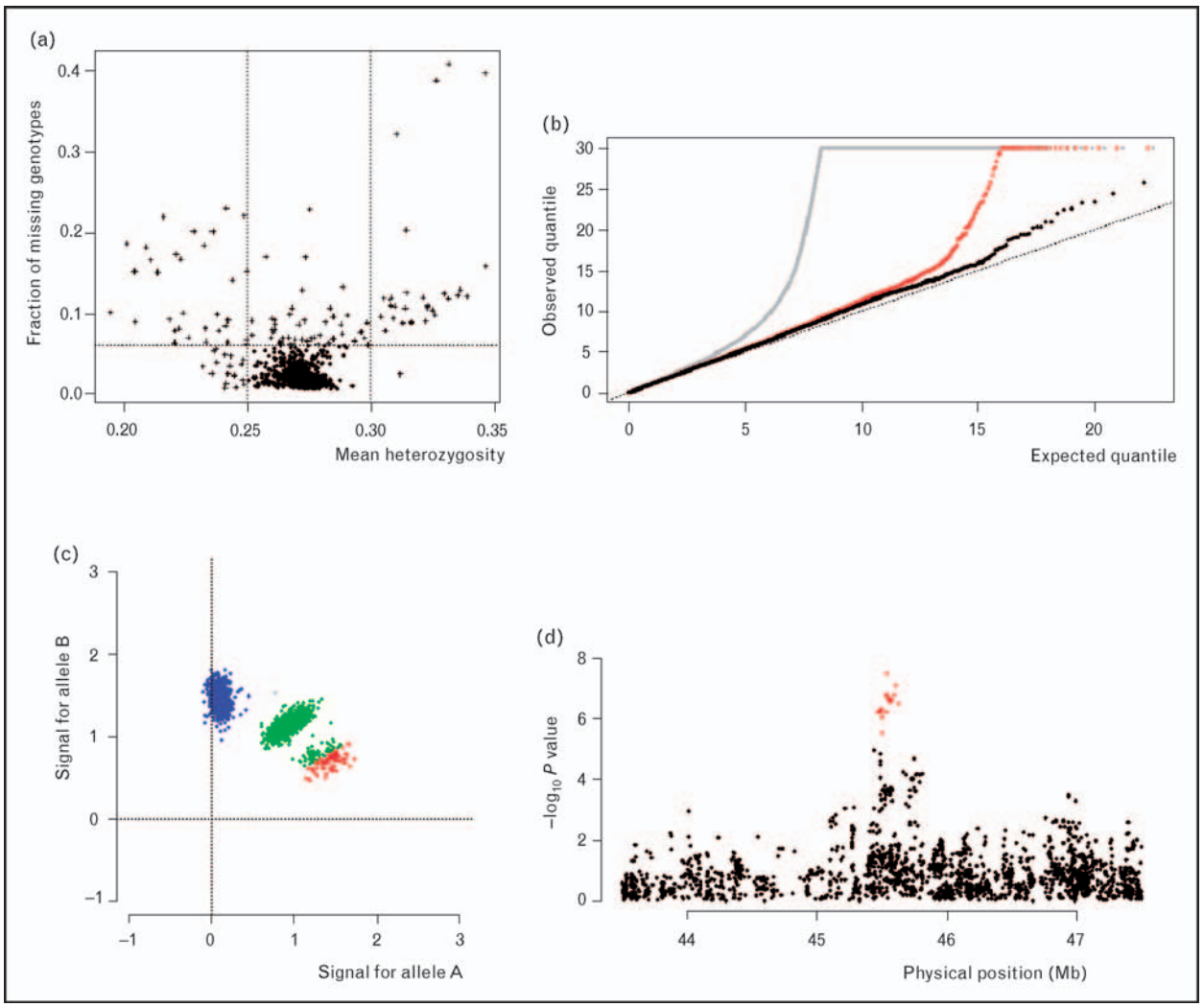
Genotyping accuracy

In theory, a researcher can attain perfect call rates by assigning genotypes to the most likely call regardless of the quality metric. Thus it has become necessary in GWAS to provide an indication of genotyping accuracy in addition to reporting the extent of missing genotypes. This can be achieved by measuring the amount of concordant genotypes between either sample duplicates or SNPs in perfect linkage disequilibrium (with pairwise $r^2 = 1$ in all the HapMap panels). The latter strategy assumes that SNPs found in perfect linkage disequilibrium across all four HapMap panels are almost certain to be in perfect linkage disequilibrium in the study population.

Clusterplot checking

It is important to check the fidelity of the genotype assignment for SNPs which exhibit evidence of putative trait association, since erroneous genotyping can often introduce unwarranted association signals [58]. When

Figure 2 Graphical displays of quality control statistics



(a) A plot of the fraction of missing genotype against the mean heterozygosity for each sample out of 1500 individuals, assessed across 490 032 autosomal single nucleotide polymorphisms (SNPs). The fraction of missing genotype is calculated as the extent of null genotype calls for each sample out of 490 032 calls, and the mean heterozygosity is calculated as the proportion of heterozygous genotypes out of 490 032. Only samples with less than 6% missing genotypes (horizontal dotted lines) and mean heterozygosity of between 0.25 and 0.30 (vertical dotted lines) are included for further analyses. (b) A quantile–quantile plot for the Armitage trend test statistic for the same data, in which observed values greater than 30 have been scaled to 30 for presentation purposes. Grey circles represent the data before filtering; red circles represent the data after removing SNPs with over 5% missing genotypes; black circles represent the data after further removing SNPs with Hardy–Weinberg equilibrium test statistic greater than 28 (approximately corresponding to a P value of 10^{-7}) and minor allele frequency less than 1%. The dotted line indicates the expected distribution of the test statistic. (c) A clusterplot of a SNP with erroneous genotype calling. The color of each dot represents the assigned genotype, subject to satisfying the threshold for the confidence metric. Thus erroneous genotypes may occur even at stringent thresholds. (d) A signal plot across a small region on a chromosome, where the x-axis represents the physical position in megabases, and the y-axis shows the $-\log_{10} P$ value for the Armitage trend test. Each circle represents a SNP on the genotyping platform, and circles in red represent SNPs with significances less than 10^{-5}. As genotyping errors can introduce anomalous association signals, a clustering of SNPs with suggestive evidence of trait association rules out the possibility of false positives attributed to genotyping errors.

the genotypes for case and control cohorts are called independently, it is possible the genotypes at a SNP may be correctly assigned for one cohort but erroneously assigned for the other (Fig. 2c). This can lead to an overrepresentation of a particular genotype in one cohort (the heterozygous genotype in our example in Fig. 2c), thus producing a spuriously large association signal. Consistently strong signals of association across a collection of SNPs in close proximity (thus more likely to be in high linkage disequilibrium) can help to rule out serious genotyping errors (Fig. 2d), although this may not be possible in the absence of a series of nearby hits or when tagging SNPs are used. Visual inspection of the clusterplot for each trait-associated SNP is still the recommended strategy for ascertaining the accuracy of the genotyping [13••].

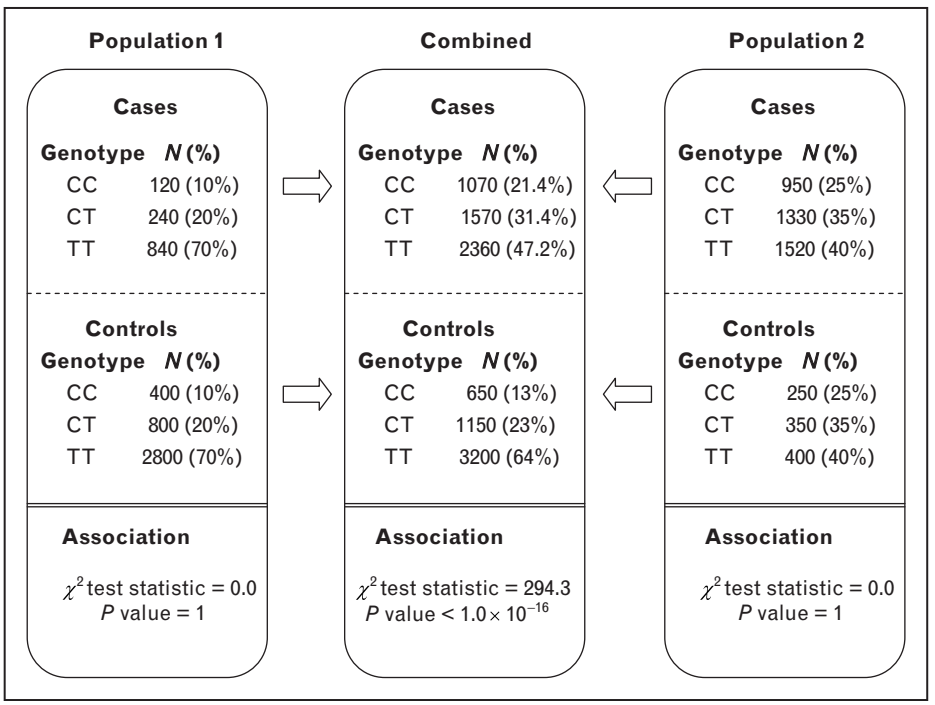
Additional single nucleotide polymorphism-level quality control

As denser genotyping platforms become available and increasingly more markers are assayed simultaneously, there will be a greater reliance on automated procedures for evaluating the quality of the genotyping. This is important for downstream analyses such as haplotype phasing and identifying population structure, and for comparisons against publicly available datasets with high-quality genotypes like the HapMap database [3,4••], since inaccurate genotyping can introduce misleading differences. A number of metrics have been introduced as part of the procedures for assigning genotypes and these provide measures of genotyping quality for the entire SNP in addition to the confidence measure for the genotype assigned to each sample [15•,55,60]. As standard quality control criteria typically rely on extreme behaviors, including gross departures to HWE, low minor allele frequencies and high rates of missingness, problematic SNPs which do not present extreme statistics for these quality control criteria may not actually be detected. SNP-level metrics, when used with standard filters, have been shown to be highly effective in identifying SNPs with problematic genotyping [61].

Population structure

Population structure refers to the genetic differences that exist between individuals from different groups, populations or geographical regions. As the effects of population structure usually result in differences in the allele frequencies of genetic variants between populations, undetected or unaccounted population structure in an association study with unrelated individuals has the potential to result in confounding and biases [11••,32–34,62–69] (Fig. 3). This is particularly relevant in studies of complex diseases since the magnitude of association signals from each of multiple disease genes may be marginal, to the extent that they are comparable to or even dwarfed by confounding signals from unaccounted population structure. The increasing sample sizes recommended for GWAS also meant that such studies are increasingly susceptible to confounding from finer levels of population differences. Assessing the presence of population structure in GWAS has thus become a permanent feature on the analytical

Figure 3 A representation of how differences in genotypic (or allelic) frequencies across different populations can introduce false signals of association



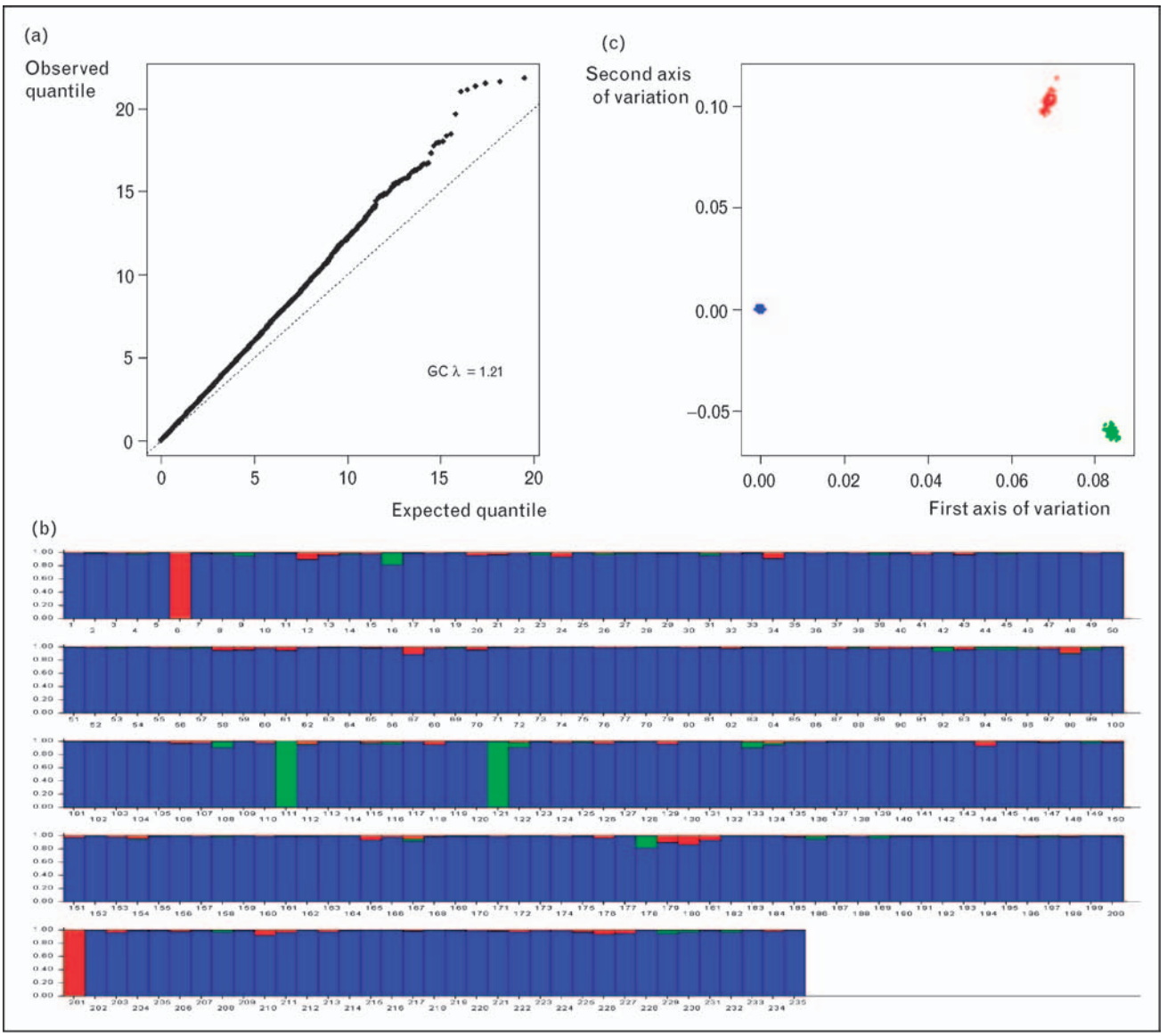
In this artificial example with 5000 cases and 5000 controls, samples were recruited from two populations. Within each population, the frequencies for the three genotypes are identical between the affected and unaffected samples, resulting in no evidence of association when the case–control data are analyzed within each population. If data from the two populations were merged without acknowledging population differences, the resultant case–control data can inflate the test statistic and produce a spurious association signal. The appropriate approach here will be to pool the data using metaanalysis procedures (i.e. with a Cochran–Mantel–Haenszel test), which will correctly yield a nonsignificant result.

checklist for GWAS to guard against false positive associations introduced by population differences.

There are a number of established statistical strategies for detecting population structure, of which those commonly used in genome-wide studies include the following: genomic control, which estimates the degree of inflation of the test statistic but necessarily assumes that existing population structure has a uniform influence throughout the entire genome [30,70] (Fig. 4a); structure, which assumes (and infers) a fixed number of possible ancestral backgrounds that every sample either partially or wholly belongs to [5,7] (Fig. 4b); and principal component analysis with Eigenstrat, which infers informative axes of genetic variation and locates each sample on this high-dimensional map of continuous genomic variation [11••] (Fig. 4c).

These strategies typically rely on having a dense set of trait-independent and unlinked genetic data. As most well designed genotyping platforms prioritize the use of tagging

Figure 4 Graphical representation of population structure



(a) A quantile–quantile plot of the Armitage trend test statistic for a genome-wide association study in a sample with population structure, with the expected distribution indicated by the dotted line. The estimated genomic control (GC) inflation factor (λ) was 1.21, representing the χ^2 test statistic for each single nucleotide polymorphism has been inflated by 21% due to population differences. (b) A graphical representation of the structure-assigned subpopulation membership for each sample. In this example with 235 samples, structure was run assuming three subpopulations (as represented by the colors blue, green and red) and samples with multiple colors are admixed for the respective populations. (c) A plot of the first two axes of genetic variation inferred using the program Eigenstrat for the 269 HapMap samples. The first axis (*x*-axis) effectively separates the YRI from the CHB + JPT and CEU, while jointly the two axes separate the three HapMap panels.

SNPs, identifying a set of unlinked SNPs for detecting population structure is straightforward. For platforms that do not explicitly rely on tagging SNPs (for example, the Affymetrix 500K array), a simple strategy will be to utilize any SNPs subject to the condition that the inter-marker distance is greater than 200 kb [3,37••]. In our experience, thinning the collection of SNPs by using one in every five consecutive SNPs on the array can minimize linkage disequilibrium between the SNPs used. A greater challenge lies in ensuring that SNPs used are independent of the trait of interest, since assessing population structure typically happens prior to association testing, and there is no way to tell *a priori* whether a SNP is independent of the trait. One common strategy is to avoid genomic regions with reported associations (for example, the major histocompatibility complex region for autoimmune diseases), although this relies on having comprehensive and reliable genetic catalogues of the disease. We advocate a simple solution to minimize the likelihood of including associated SNPs: perform a round of association analysis assuming no population structure exists, remove all SNPs with at least marginal evidence of phenotypic association and use the remaining markers as the superset for identifying SNPs to use for identifying population structure.

Common strategies to account for the presence of population structure in association studies include the following: dividing the test statistic at every SNP by the inflation factor estimated from genomic control [69]; performing stratified analyses based on the assigned subpopulation membership from structure [71,72] at every SNP; and using the coordinates from the axes of genetic variation inferred from eigen analysis as covariates in a regression framework [11••].

While there is ample literature on the validity of these approaches, the relative merits and shortcomings of each method need to be reevaluated in the current era of GWAS. Genomic control requires almost no additional analysis and is easily implemented after testing for associations. Assuming population differences exert a uniform influence across the genome, however, ignores the fact that the genetic architecture of each individual is a complicated mosaic from different genetic backgrounds. Simply dividing the test statistic by an estimated inflation factor is naïve and potentially misleading. While highly sophisticated and accurate, the present version of structure is slow to run on large datasets. This limits the application of structure in GWAS to small subsets of SNPs (typically less than 20 000), potentially neglecting the information from a large fraction of the available data. Initially designed with GWAS in mind, principal component analysis appears to be the preferred method for handling population structure in large genetic studies. The approach handles a large number of SNPs across thousands of samples effortlessly, is relatively fast to implement, and the use of inferred principal components as covariates in a regression analysis is both intuitive and appealing.

As the sample sizes increase in GWAS and denser genotyping platforms become available, it is expected that sophisticated statistical methods will be developed to detect and handle ever finer levels of population differences between individuals. One area in need of methodological development is the management of population structure in replication studies. These experiments typically assay a handful of putative trait-associated SNPs on additional samples to verify the detected associations. It is a challenge to perform accurate inference of any population structure in the new samples given the paucity of the available SNP data. While typing additional ancestry-informative markers in a replication study is a possible alternative, this is seldom performed for economic reasons. Well crafted methodological approaches bridging the inference of population structure for samples in the main GWAS and subsequent replication experiments will be useful as these studies are expanded to populations with diverse genetic backgrounds.

Conclusions

Studies on the genomics of common diseases and complex traits have successfully uncovered a number of novel trait-associated genetic variants. This is set to continue as the genome-wide approach is extended across an ever-increasing spectrum of diseases. While researchers in population genetics are learning to cope with some of the statistical challenges in genome-wide studies, there remain a number of analytical obstacles in the marriage of genetic and epidemiological data for understanding gene–gene and gene–environment interactions [73]. These obstacles need to be surmounted before the complex interplay between genetic and environmental factors can be fully understood, thereby truly achieving the full potential of a GWAS.

Acknowledgements

The author acknowledges support from the Grand Challenges in Global Health and the Wellcome Trust. The author also thanks K.S. Small and X.L. Sim for their helpful comments in improving the article.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

Additional references related to this topic can also be found in the Current World Literature section in this issue (pp. 169–170).

1 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409:860–921.

2 Venter JC, Adams MD, Myers EW, *et al.* The sequence of the human genome. Science 2001; 291:1304–1351.

3 International HapMap Consortium. A haplotype map of the human genome. Nature 2005; 437:1299–1320.

4 •• International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007; 449:851–861.
This report describes the findings from the second phase of the International HapMap project over 3.1 million SNPs, and provides a detailed discussion on the coverage of current genotyping platforms. In addition, the report discusses the structure of linkage disequilibrium and investigates the extent of recombination in the human genome.

5 Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000; 155:945–959.

6 Stephens M, Smith N, Donnelly P. A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 2001; 68:978–989.

7 Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 2003; 164:1567–1587.

8 de Bakker PI, Yelensky R, Pe'er I, *et al.* Efficiency and power in genetic association studies. Nat Genet 2005; 37:1217–1223.

9 • Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 2006; 78:629–644.
This article describes an extremely fast method for inferring the haplotypic phase from genotype data and has the ability to infer missing genotypes.

10 Marchini J, Cutler D, Patterson N, *et al.* A comparison of phasing algorithms for trios and unrelated individuals. Am J Hum Genet 2006; 78:437–450.

11 •• Price AL, Patterson NJ, Plenge RM, *et al.* Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006; 38:904–909.
This article describes a procedure for detecting population structure in genetic data and also introduces a method for correcting the presence of population structure in an association study.

12 • Clayton D, Leung HT. An R package for analysis of whole-genome association studies. Hum Hered 2007; 64:45–51.
This article describes a user-friendly computational package for analyzing data from GWASs.

13 •• Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3000 shared controls. Nature 2007; 447:661–678.
This landmark study describes in detail a number of methodological issues related to GWASs and proposes a number of sophisticated solutions to these issues. This study also reports on seven GWASs for seven common human diseases conducted across a total of 17 000 individuals.

14 •• Marchini J, Howie B, Myers S, *et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 2007; 39:906–913.
This article describes a sophisticated procedure to statistically infer genotypic information for all the SNPs in the International HapMap project from data in a genome-wide study, in essence performing in-silico fine mapping.

15 • Teo YY, Inouye M, Small KS, *et al.* A genotype calling algorithm for the Illumina BeadArray platform. Bioinformatics 2007; 23:2741–2746.
This article describes a calling algorithm for the Illumina BeadArray platform which, in addition to the usual genomic DNA, can also call genotypes for hybridization data from whole-genome amplified DNA.

16 Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 2005; 6:95–108.

17 Wang WY, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nat Rev Genet 2005; 6:109–118.

18 Todd JA, Walker NM, Cooper JD, *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. Nat Genet 2007; 39:857–864.

19 Diabetes Genetics Institute. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007; 316:1331–1336.

20 Zeggini E, Weedon MN, Lindgren CM, *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007; 316:1336–1341.

21 Scott LJ, Mohlke KL, Bonnycastle LL, *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 2007; 316:1341–1345.

22 Hunter DJ, Kraft P, Jacobs KB, *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. Nat Genet 2007; 39:870–874.

23 Stacey SN, Manolescu A, Sulem P, *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. Nat Genet 2007; 39:865–869.

24 van Heel DA, Franke L, Hunt KA, *et al.* A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL2I. Nat Genet 2007; 39:827–829.

25 Dina C, Meyre D, Gallina S, *et al.* Variation in FTO contributes to childhood obesity and severe adult obesity. Nat Genet 2007; 39:724–726.

26 Frayling TM, Timpson NJ, Weedon MN, *et al.* A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. Science 2007; 316:889–894.

27 Evangelou E, Trikalinos TA, Salanti G, Ioannidis JPA. Family-based versus unrelated case-control designs for genetic associations. PLoS Genet 2006; 2:1147–1155.

28 McGinnis R, Shifman S, Darvasi A. Power and efficiency of the TDT and case-control design for association scans. Beh Genet 2002; 32:135–144.

29 Cavalli-Sforza LL, Menozzi P, Piazza A. The history and geography of human genes. Princeton: Princeton University Press; 1994.

30 Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. Genet Epidemiol 2001; 20:4–16.

31 Cardon LR, Paler LJ. Population stratification and spurious allelic association. Lancet 2003; 361:598–604.

32 Freedman ML, Reich D, Penney KL, *et al.* Assessing the impact of population stratification on genetic association studies. Nat Genet 2004; 36:388–393.

33 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.

34 Helgason A, Yngvadóttir B, Hrafnkelsson B, *et al.* An Icelandic example of the impact of population structure on association studies. Nat Genet 2005; 37:90–95.

35 Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. Am J Hum Genet 2004; 74:765–769.

36 Carlson CS, Eberle MA, Rieder MJ, *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. Nat Genet 2003; 33:518–521.

37 •• de Bakker PI, Burtt NP, Graham RR, *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. Nat Genet 2006; 38:1298–1303.
This article describes how transferable tagging SNPs are in general. This has important consequences for GWASs since most genotyping platforms use tagging SNPs to increase the effective coverage, and these tagging SNPs are typically identified from the four populations in the HapMap project.

38 • Barrett JC, Cardon LR. Evaluating coverage of genome-wide association studies. Nat Genet 2006; 38:659–662.
This article discusses the coverage of the human genome conferred by the earlier genotyping technologies, and is important for the understanding of how coverage can vary between populations.

39 Carlson CS, Eberle MA, Rieder MJ, *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. Am J Hum Genet 2004; 74:106–120.

40 • Conrad DF, Jakobsson M, Coop G, *et al.* A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet 2006; 38:1251–1260.
This article reports an important finding related to the extent of haplotype diversity across multiple populations, and discusses in detail the applicability of genome-wide studies across multiple populations.

41 Chapman JM, Cooper JD, Todd JA, Clayton DG. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum Hered 2003; 56:18–31.

42 Clayton D, Chapman J, Cooper J. Use of unphased multilocus genotype data in indirect association studies. Genet Epidemiol 2004; 27:415–428.

43 Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005; 37:413–417.

44 Waldron ERB, Whittaker JC, Balding DJ. Fine mapping of disease genes via haplotype clustering. Genet Epi 2006; 30:170–179.

45 • Li Y, Abecasis GR. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet 2006; S79:2290.
This article describes a novel method for inferring both the haplotypic phase and missing data from unphased genotype data.

46 Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 2007; 3:e114.

47 Pfeiffer RM, Gail MH. Sample size calculations for population- and family-based case-control association studies on marker genotypes. Genet Epidemiol 2003; 25:136–148.

48 De La Vega FM, Gordon D, Su X, *et al.* Power and sample size calculations for genetic case/control studies using gene-centric SNP maps: application to human chromosomes 6, 21, and 22 in three populations. Hum Hered 2005; 60:43–60.

49 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.

50 Liu W, Di X, Yang G, *et al.* Algorithms for large-scale genotyping microarrays. Bioinformatics 2003; 19:2397–2403.

51 Di X, Matsuzaki H, Webster TA, *et al.* Dynamic model based algorithms for screening and genotyping over 100K SNPs on oligonucleotide microarrays. Bioinformatics 2005; 21:1958–1963.

52 Rabbee N, Speed T. A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics 2006; 1:7–12.

53 Affymetrix Inc. BRLMM: an improved genotype calling method for the GeneChip Human Mapping 500K Array Set; 2006. http://www.affymetrix.com/support/technical/whitepapers/brlmm_whitepaper.pdf. [Accessed 16 January 2007]

54 Affymetrix Inc. BRLMM-P: a genotype calling method for the SNP 5.0 array; 2007; http://www.affymetrix.com/support/technical/whitepapers/brlmmp_whitepaper.pdf. [Accessed 16 January 2007]

55 Plagnol V, Cooper JD, Todd JA, Clayton DG. A method to address differential bias in genotyping in large scale association studies. PLoS Genet 2007; 3:e74.

56 Bolstad BM, Irizarry RA, Astrand M, Speed TP. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 2003; 19:185–193.

57 Kermani BG. Artificial intelligence and global normalization methods for genotyping. US Patent 2006; 20060224529.

58 Clayton DG, Walker NM, Smyth DJ, *et al.* Population structure, differential bias and genomic control in a large-scale case-control association study. Nat Genet 2005; 37:1243–1246.

59 Teo YY, Fry AE, Clark TG, *et al.* On the usage of HWE for identifying genotyping errors. Ann Hum Genet 2007; 71:701–703.

60 Xiao Y, Segal MR, Yang YH, Yeh RF. A multiarray multi-SNP genotyping algorithm for Affymetrix SNP microarrays. Bioinformatics 2007; 23:1459–1467.

61 Teo YY, Small KS, Clark TG, Kwiatkowski DP. Perturbation analysis: a simple method for filtering SNPs with erroneous genotyping in genome-wide association studies. Ann Hum Genet (in press).

62 Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994; 265:2037–2048.

63 Altshuler D, Kruglyak L, Lander E. Genetic polymorphisms and disease. N Engl J Med 1998; 338:1626.

64 Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000; 405:847–856.

65 Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet 2001; 2:91–99.

66 Peltonen L, Palotie A, Lange K. Use of population isolates for mapping complex traits. Nat Rev Genet 2001; 1:182–190.

67 Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. Am J Hum Genet 2002; 71:304–311.

68 Thomas DC, Witte JS. Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 2002; 11:505–512.

69 Ziv E, Burchard EG. Human population structure and genetic association studies. Pharmacogenomics 2003; 4:431–441.

70 Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999; 55:997–1004.

71 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. Am J Hum Genet 2000; 67:170–181.

72 Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. Theor Popul Biol 2001; 60:227–237.

73 Manolio TA, Collins FS. Genes, environment, health, and disease: facing up to complexity. Hum Hered 2007; 63:63–66.