Biostatistics (2010), **11**, 1, *pp.* 70–81
doi:10.1093/biostatistics/kxp038
Advance Access publication on October 14, 2009

Exploratory data analysis in large-scale genetic studies

YIK Y. TEO*

Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK, Wellcome Trust Sanger Institute, Hinxton, Cambridge CB10 1SA, UK and Department of Statistics and Applied Probability and Centre for Molecular Epidemiology, National University of Singapore, Singapore 117546
teo@well.ox.ac.uk

SUMMARY

Genome-wide association studies (GWAS) have become the method of choice for investigating the genetic basis of common diseases and complex traits. The immense scale of these experiments is unprecedented, involving thousands of samples and up to a million variables. The careful execution of exploratory data analysis (EDA) prior to the actual genotype–phenotype association analysis is crucial as this identifies problematic samples and poorly assayed genetic polymorphisms that, if undetected, can compromise the outcome of the experiment. EDA of such large-scale genetic data sets thus requires specialized numerical and graphical strategies, and this article provides a review of the current exploratory tools commonly used in GWAS.

Keywords: Exploratory data analysis; Genetic association studies.

1. INTRODUCTION

Exploratory data analysis (EDA) refers to the process of summarizing a data set using informative numerical statistics, well-constructed tables, or graphical visualizations, serving to (i) assess data quality and fidelity; (ii) check the validity of specific assumptions necessary for statistical inference; and (iii) identify the appropriate analytical and statistical strategies in lieu of the properties exhibited by the data set (Tukey, 1977). The careful execution of EDA is arguably the most important process in data analysis since the undetected presence of problematic or nonrepresentative data or the use of statistical tools that are inappropriate for the data set can result in meaningless and often wrong conclusions, jeopardizing an otherwise perfectly designed research experiment.

In the field of genetic epidemiology, technological advances coupled with the availability of extensive databases on human genetic variations have allowed the systematic survey of the human genome for genetic variants which are associated with the disease or trait phenotypes (International HapMap Consortium, 2007), employing an experimental procedure known as a genome-wide association study (GWAS). This experimental design recently became the method of choice for investigating the genetic etiology of common diseases and complex traits, which is expected to depend on a complex interplay between

*To whom correspondence should be addressed.

environmental factors and multiple genetic variants each contributing a modest effect (Hirschhorn and Daly, 2005; Wang *and others*, 2005). In order to avoid the issue of multiplicity from testing potentially a million genetic variants, known as single nucleotide polymorphisms (SNPs), the scientific community has adopted strict definitions of statistical significance (e.g. $p < 5 \times 10^{-7}$, NCI-NHGRI Working Group on Replication in Association Studies, 2007) which, together with the expected modest effect sizes, dictate the need for large sample sizes typically involving thousands of subjects. Such large-scale genetic studies offer unprecedented challenges in assessing data quality and fidelity, especially since data of suspect quality can introduce statistical artefacts that either mimic the marginal effect sizes often associated with genuine biological findings or mask true associations. Well-designed EDA tools are thus meant to identify problematic data while avoiding the excessive removal of samples and SNPs, which lowers statistical power by reducing both the sample size and the effective genomic coverage with a sparser SNP set.

While the aims of EDA in GWAS are similarly to identify problems or latent structure in the data, the actual EDA methods in GWAS are very different from conventional exploratory tools like histograms, boxplots, or 5-number summaries. This paper aims to introduce the specialized EDA tools used in GWAS and discuss their relevance in affecting the nature of downstream association analysis. Multiple layers of data manipulation need to happen prior to the actual genotype–phenotype association analysis. We provide a brief review of the process of genotype calling in Section 2. Section 3 introduces exploratory techniques necessary for assessing SNP data quality, while Section 4 focuses on methods for sample quality control. Section 5 provides a discussion on the EDA techniques used for identifying population structure. Section 6 provides a brief exposition on the use of graphical visualizations to investigate the research hypothesis and interpret the research findings in the context of localizing the actual disease-causing variants.

2. Source of missingness and errors

Understanding the source of missing and erroneous data is important in designing suitable tools for data exploration. This is especially important if the amount of missing data correlates with the extent of errors and has the potential to introduce systematic biases in the analysis. The response variable in a genetic association study is either dichotomous categorical in a case–control setting or numerical when investigating a quantitative trait like height or weight. It is often assumed that there is no error or missingness in the response phenotype and data issues fundamentally stem from the explanatory variables—the SNPs. Therefore, understanding how the data at each SNP (i.e. the genotypes) is generated is useful in addressing the issue of problematic data in a genetic study.

The process of genotyping an SNP involves the assessment of fluorescence intensities for the possible alleles from probes which target a specific genomic region. Most genome-wide studies rely on commercial genotyping arrays from either Affymetrix (Santa Clara, CA, USA) or Illumina (San Diego, CA, USA) that allow up to a million SNPs to be assayed simultaneously. Translating the hybridization intensities of these SNPs into actual genotypes requires the use of automated calling procedures which essentially compare the relative intensity of the fluorescence signals between the possible alleles to decide on the likely genotype (Carvalho *and others*, 2007; Teo, Inouye, *and others*, 2007; Wellcome Trust Case Control Consortium, 2007; Korn *and others*, 2008; see Figure 1). These unsupervised calling algorithms evaluate the intensity profiles from multiple samples concurrently and, on the basis of the characteristics and positions of the genotype clouds, calculate the posterior probabilities of the 3 possible genotypes for every sample at each SNP. The genotype yielding the highest probability is assigned if this probability exceeds a designated threshold, otherwise a missing outcome is assigned (Figure 1(a)). Misinterpretation of the intensity profiles for ambiguous or unusual genotype clusters can result in a poor calibration of the cluster characteristics, yielding unwarranted confidence in genotype assignments which produce

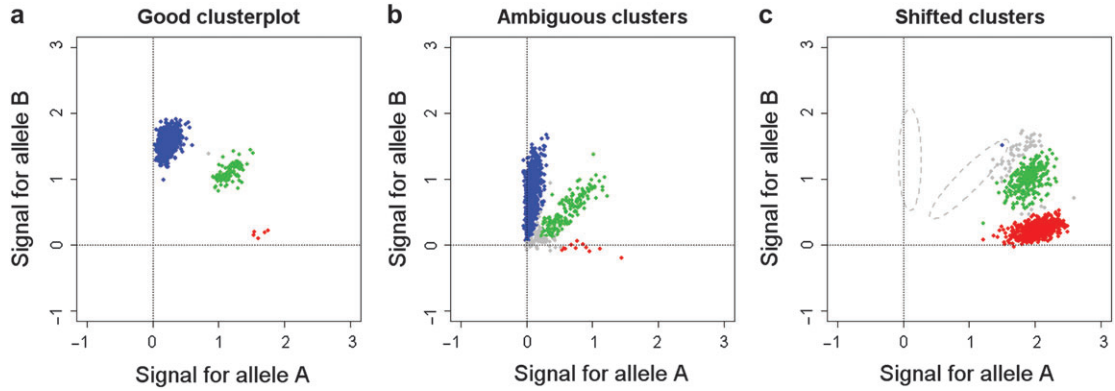

Fig. 1. Clusterplot representations of the hybridization intensities. Each figure represents the intensity data for all the samples at an SNP, where each solid circle represents the intensity profile for a sample and the axes measure the hybridization signals for the 2 possible alleles, generically designated as "A" and "B" here. The genotype calls from a well-calibrated automated calling algorithm (Wellcome Trust Case Control Consortium, 2007) are superimposed as the different colours, with red corresponding to an "AA" genotype, green for an "AB" genotype, blue for a "BB" genotype, and grey to represent uncalled or missing genotype. (a) An SNP with distinct clusters and accurately assigned genotypes; (b) An SNP with overlapping clusters, where a well-calibrated calling algorithm will assign missing genotypes to samples in the overlapping region due to ambiguity; and (c) An SNP where the clusters for genotypes "AB" and "BB" are shifted relative to the usual positions as represented by the dashed ellipses in grey, resulting in a high degree of missing and erroneous genotype assignments.

erroneous calls that have no bearing to the actual genotypes (Figures 1(b) and (c)).This process of genotype calling using automated procedures can thus introduce missing and erroneous data, which may correlate across unrelated samples, given the reliance on cluster characteristics inferred from multiple samples.

3. SNP QUALITY CONTROL

The fidelity of the genotype assignment for every SNP should ideally be assessed prior to the actual association analysis, confining the search only to SNPs with accurate genotyping. The preferred strategy for checking genotyping accuracy is to visually assess the clustering via a "clusterplot" (Wellcome Trust Case Control Consortium, 2007), which superimposes the genotype calls onto the intensity profiles from all samples at an SNP (see Figure 1). However, the unprecedented number of SNPs investigated in a GWAS makes it impossible to visually evaluate the integrity of the genotype data for every SNP. Instead, this exploratory process has been deferred, and only the clusterplots for SNPs with suggestive signals of phenotypic association are checked. This however means that downstream analysis investigating the extent of population structure (see later), haplotype phasing, and analysis of natural selection is likely to involve SNPs which contain missing and erroneous genotypes. As such analyses often compare the existing data against reference data sets with highly accurate genotypes (e.g. from the International HapMap Project), the inclusion of SNPs with problematic genotyping, however subtle, can artificially introduce noise in the comparisons, suggesting greater differences between the data sets than actually present. It is thus common to explore a number of summary metrics that are informative for identifying SNPs with poor quality genotyping.

3.1 *Hardy–Weinberg equlibrium*

An SNP is mathematically defined to be in a state of Hardy–Weinberg equilibrium (HWE) if the probability of observing a particular genotype is equal to the probability of observing the 2 alleles independently. Testing for departure from HWE involves comparing the observed genotype frequencies to the expected counts inferred from the allele frequencies, often via a Pearson's chi-squared test or a Fisher's exact test, and is meant to reflect nonrandom mating or genetic drift. Checking the distribution of the genotypes at an SNP for adherence to HWE has conventionally been used to identify sampling biases, particularly amongt the controls in a case–control experiment. Given the number of SNPs analyzed in a typical GWAS and the associated problem with multiple testing, strict conformity to HWE has become less relevant. Instead, it is common to test for gross departures from HWE using liberal significance thresholds (i.e. $P_{\mathrm{HWE}} < 10^{-4}$ to 10^{-7}) since automated genotype calling procedures often introduce errors which produce genotype distributions that significantly violate HWE (Figure 1(c); Teo, Fry, *and others*, 2007).

3.2 *SNP missingness*

The extent of missing genotypes from well-calibrated calling algorithms is often a useful surrogate for the accuracy of the genotyping since high missingness indicates either noisy intensity profiles resulting in overlapping or ambiguous genotype clouds (Figure 1(b)) or atypical cluster characteristics that the genotype calling algorithm is not confident in interpreting (Figure 1(c)). Removing SNPs with high levels of missing genotypes reduces the likelihood that an SNP with dubious genotyping is retained for downstream analyses, and a common standard is to exclude an SNP from further analyses if there is more than 5% missing data (Wellcome Trust Case Control Consortium, 2007). Additionally, in a case–control experiment, it is important to check that SNPs with putative disease association do not contain significant differences in the level of missingness between cases and controls since such differential bias has the potential to produce spurious association (Clayton *and others*, 2005).

3.3 *Minor allele frequency*

A genetic variant with a low minor allele frequency results in low genotype counts for at least 1 of the 3 possible genotypes, yielding sparse genotype clouds which can confound the calling process. It is common to exclude SNPs with minor allele frequencies that are less than 1% as they are more likely to be affected by genotyping errors and are generally less informative and underpowered for an association study.

3.4 *Perturbation analysis*

Perturbation analysis is a diagnostic tool specifically designed to assess the quality of the genotyping at the SNP level (compared to the posterior probabilities from calling algorithms which reflect the quality of the genotypes for individual samples at an SNP) and checks the stability of the calls to minor perturbation of the hybridization intensities (Teo, Small, *and others*, 2008). By applying another round of genotype calling to intensity data where white noise of a suitably small magnitude has been introduced, the amount of discordant genotypes between the calls made before and after perturbing the intensities can be evaluated. This yields a metric which is useful for identifying SNPs with problematic genotyping, particularly for SNPs with ambiguous or overlapping genotype clusters. As this summary metric for data quality relies on the algorithm used for assigning genotypes, it is commonly implemented in genotype calling software (Plagnol *and others*, 2007; Teo, Inouye, *and others*, 2007), removing SNPs with greater than 5% discordant genotypes.

4. SAMPLE CONTROL

The quality of the genotype data for each sample fundamentally relies on the quality of the input DNA, and insufficient or degraded DNA can lead to hybridization failures, introducing a greater extent of missing data across the assayed SNPs. However, assessing data integrity for each sample requires additional exploratory tools since sample handling errors and the inclusion of related samples are not detectable via the extent of missing genotypes for each sample.

4.1 *Sample missingness*

The proportion of missing genotypes for each sample across all the SNPs assayed is useful for identifying poorly genotyped samples since the greater extent of missingness suggests mechanical failure due to faulty genotyping chips or poor quality DNA. While it is common to remove samples if they contain more than 5% missing data (Wellcome Trust Case Control Consortium, 2007), it is often more appropriate to determine the missingness threshold after visualizing the distribution of missing data proportions in the study (Figure 2(a)) as particular laboratory procedures (e.g. whole genome amplification) have been established to result in greater missingness in large-scale genotyping (Teo, Inouye, *and others*, 2008). As these laboratory procedures generally are employed on all the samples in a study and can thus result in a study-wide increase in sample missingness, the strict adherence to generic thresholds without consideration of the data at hand can dramatically reduce the number of samples for subsequent association analysis, decreasing statistical power. The caveat of including samples with higher rates of missingness is the possibility of increased false-positive associations due to problematic genotyping. However, as the clusterplots for any putative associations are expected to undergo rigorous visual assessment, the trade off for including potentially noisy samples, and thus conserving power, is simply that more clusterplots

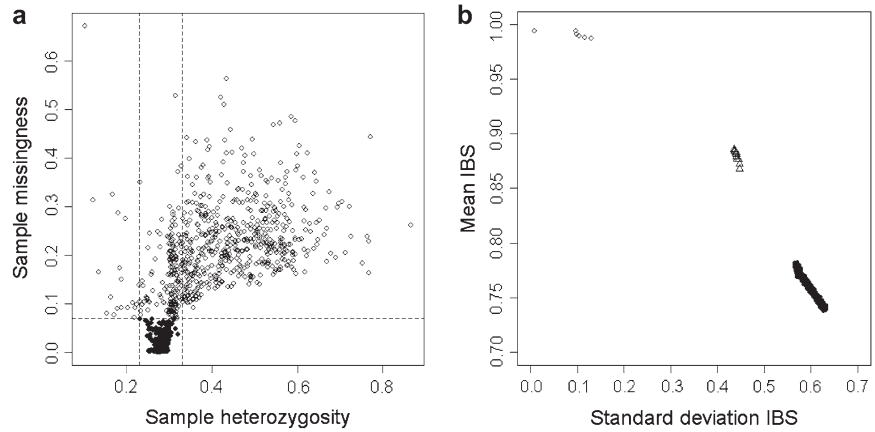

Fig. 2. Graphical exploratory tools for sample quality control. (a) The figure plots the proportion of missing genotypes for each sample against the observed heterozygosity. Each solid circle represents a sample which passes quality control and is preserved for downstream analyses, while each open circle represents a sample with either excessive level of missingness or heterozygosity and is thus removed. The dashed lines represent the designated thresholds for this data set. (b) The figure investigates the extent of sample relatedness by the amount of IBS genotypes between pairs of samples. Solid circles refer to the levels of IBS for the bulk of the pairwise comparisons between unrelated samples in this data set; open triangles indicate levels of IBS between related samples; open circles indicate monozygotic twin relationships, which often indicates the presence of duplicated samples in the experiment.

need to be checked. This is an acceptable compromise compared to the alternative of missing genuine biological signals in an underpowered study that results from excessive sample removal.

4.2 *Heterozygosity*

One method to identify sample handling errors is to assess the heterozygosity of each sample since DNA contamination often results in the generation of more heterozygous genotypes in each sample than expected. The heterozygosity of a sample in this instance is calculated as the proportion of heterozygous genotypes across all autosomal SNPs. As this summary metric can be affected by the SNP content of the genotyping array (e.g. lower heterozygosity is observed when using an array with a greater proportion of SNPs with low minor allele frequencies), the distribution of heterozygosity values across all the samples need to be considered in order to identify potential outliers. Contaminated and poor quality samples often present higher rates of heterozygosity in conjunction with a greater extent of missing data, and it is thus common to represent both sample missingness and heterozygosity in the same figure to determine the appropriate criteria for sample removal (Figure 2(a)).

4.3 *Identity-by-state*

Genetic association studies rely on detecting a greater extent of genetic similarity between individuals with the same phenotype compared to individuals across different phenotypes, and the inclusion of related samples can artificially confound this definition of genetic similarity. Sample mishandling can also result in the DNA of the same individual to be genotyped twice, resulting in duplication. By calculating the extent of allele sharing, defined as the identity-by-state (IBS), across all the autosomal SNPs between every possible pair of individuals in the study, any unexpected inclusion of related or duplicated samples can be identified. As IBS is similarly affected by the SNP content of the microarray, with a greater extent of allele sharing in the presence of more SNPs with low minor allele frequencies, it is common to visualize the distribution of IBS values using a scatterplot of the mean IBS against the standard deviation in order to detect related pairs of individuals (Figure 2(b)).

5. Population structure

Population structure is commonly defined on the basis of variations in SNP allele frequencies between populations (Wright, 1943; Balding and Nichols, 1995). Exploring the extent of population structure in any genetic study is crucial since any observed association signals from a data set with unaccounted population structure may be attributed entirely to allele frequency differences stemming from population differences (Marchini *and others*, 2004; Price *and others* 2006). The availability of hundreds of thousands of genetic markers for detecting population structure in GWAS often means that subtle interpopulation genetic variations can be identified. This is important since the statistical approach to perform the genotype–phenotype analysis depends on the extent of population structure present. Two approaches are commonly used to explore the extent of population structure in a GWAS data set: the first, using the program "eigenstrat", performs a principal components analysis on the genotype data to identify axes of variation that map every sample on a continuous spectrum of genetic variation (Price *and others*, 2006) and the second, termed "genomic control", assesses the distribution of the association test statistic to estimate the extent of overinflation as a result of population structure (Devlin and Roeder, 1999).

5.1 *Eigenstrat*

Eigenstrat calculates a correlation matrix that represents the relationship between every possible pair of samples and performs an eigen analysis on this matrix to identify the extent of clustering between the

samples (Price *and others*, 2006). The presence of any systematic genetic homogeneity in the data can be qualitatively explored using biplots or scatterplots of pairs of the obtained eigenvectors (Figure 3). A plot with randomly distributed datapoints is indicative of the absence of any systematic heterogeneity caused by population structure (Figure 3(a)); conversely the presence of clustering (Figure 3(b)) or non-random distribution of datapoints (Figure 3(c)) indicates a differential degree of homogeneity between the

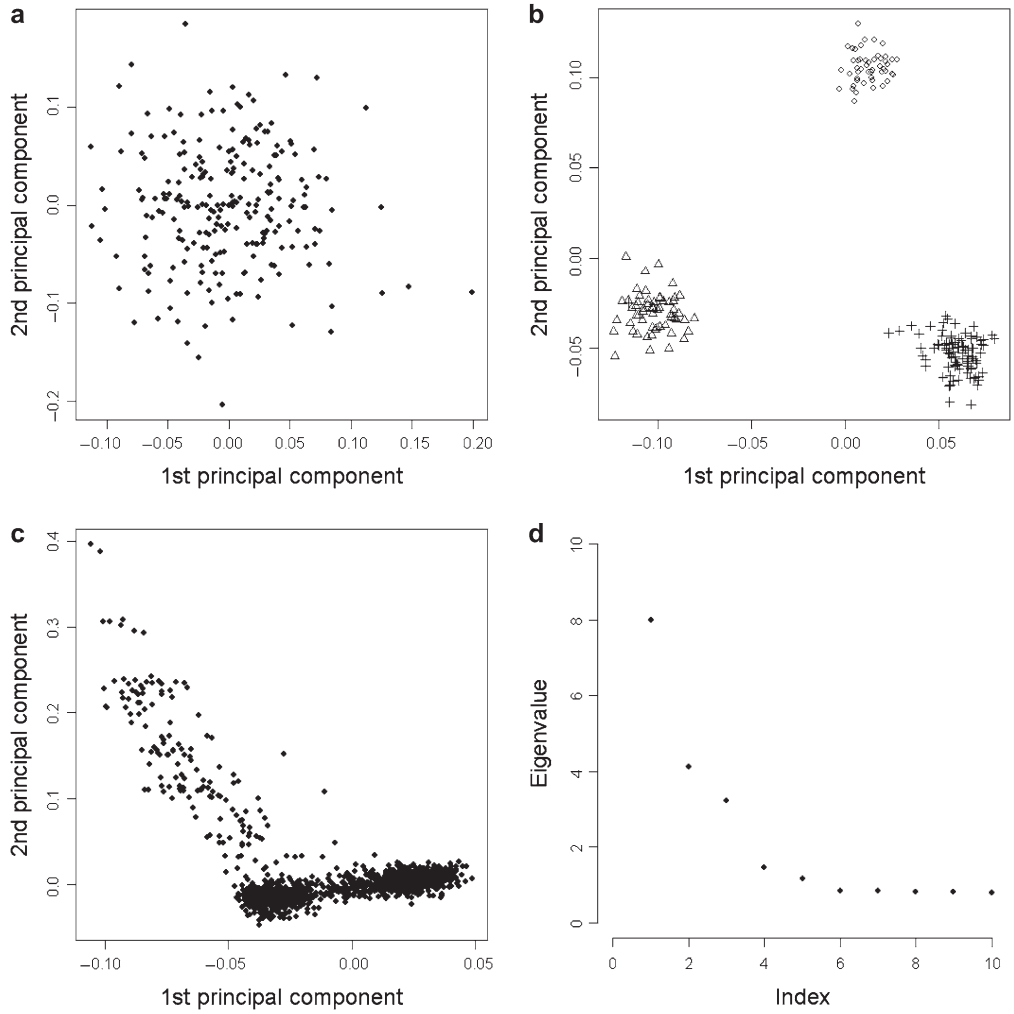

Fig. 3. Graphical exploratory analysis for population structure. (a)–(c) plot the first 2 principal components from eigen analyses of 3 separate data sets, where (a) corresponds to a data set with no population differentiation where a straightforward χ^2-test or Fisher's exact test will suffice in a case–control association analysis; (b) The figure corresponds to a data set with 3 distinct subpopulations, and the association analysis needs to be stratified according to the clustering as represented by the different symbols and subsequently combined via a Mantel–Haenszel framework; (c) The figure corresponds to a data set with substantial admixture between subpopulations, and the association analysis needs to incorporate the appropriate number of principal components as covariates in a logistic regression framework; and (d) The figure plots the values of the eigenvectors, which can be useful for identifying the number of principal components to include as covariates in the regression analysis to account for population structure, by locating the point of inflexion.

samples that necessarily has to be accounted for in any downstream association analyses. The statistical method for performing the association analysis depends crucially on the type of population structure that exists in the data since (1) a straightforward χ^2-test (and the Fisher's exact test) will suffice in a case–control association analysis without any substantial population structure as represented in Figure 3(a); (2) a stratified approach using the Mantel–Haenszel technique is required when distinct subpopulations exist in the data as represented in Figure 3(b) (Pritchard *and others*, 2000), and (3) a logistic regression incorporating the appropriate number of principal components from the eigen analysis is required to correct for the presence of admixed population structure as seen in Figure 3(c) (Price *and others*, 2006). In the last scenario, the number of principal components to include as covariates is often decided upon visualizing biplots of subsequent principal components, although this can also be decided by either plotting the decay of the eigenvalues (Figure 3(d)) or quantifying the statistical significance of the eigenvalues using a Tracy–Widom distribution (Patterson *and others*, 2006).

5.2 *Genomic control*

The presence of unaccounted population structure in an association analysis artificially introduces allele frequency differences between cases and controls which in turn inflates the statistical evidence. Genomic control refers to the procedure of estimating the degree of inflation for the association test statistic by comparing the median of the observed test statistics against the expected value from the appropriate χ^2 distribution (Devlin and Roeder, 1999). While the calculation of the inflation factor happens after the association analysis is performed and thus not technically considered exploratory, the validity and reliability of the association analysis is critically dependent on this summary statistic to uncover the presence of any systematic biases. A quantile–quantile plot (QQ-plot) of the observed test statistic against the expected distribution is widely produced in conjunction with the estimation of the inflation factor to visually represent any systematic biases (Figure 4), by checking the degree of deviation from the line of equality.

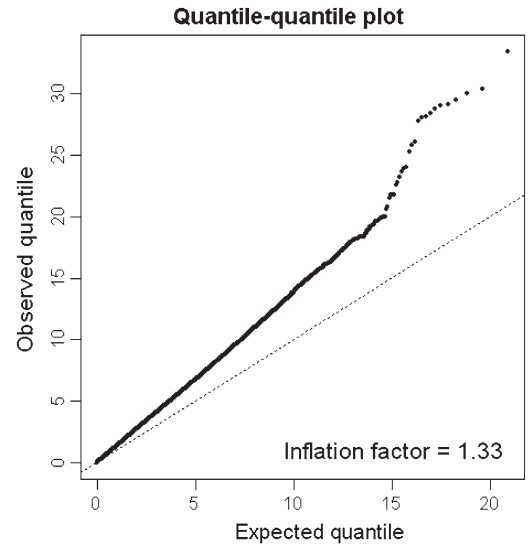

Fig. 4. QQ-plot of the Armitage-trend test statistic. Each solid circle represents the test statistic from an SNP, and the dotted line indicates the line of equality. The deviation of the circles from the line implies the existence of systematic biases, and is often an indication of the presence of population structure. The inflation factor reflects that the median test statistic is 33% greater than expected, and is used as a measure of the extent of systematic bias present.

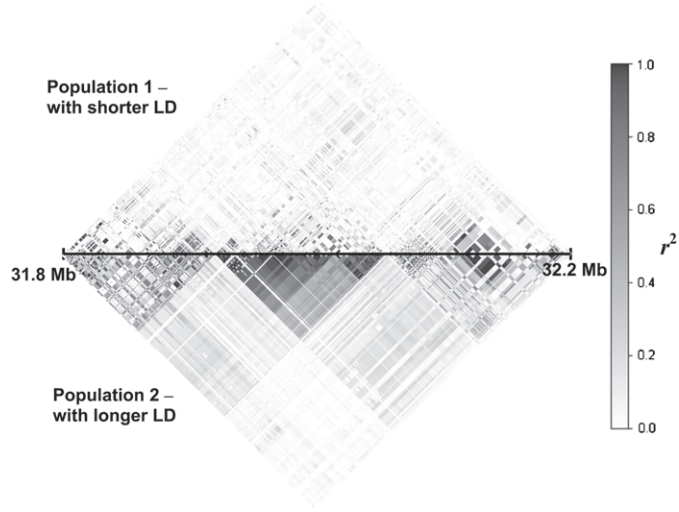

Fig. 5. Heatmap representation of LD between pairs of SNPs in a region for 2 populations. Regions in increasing intensity of red indicate increasing levels of LD, measured by the genetic correlation r^2 and as indicated by the colour key on the right. The top panel shows the regional LD for a 0.4 Mb region in a population with low and short LD, while the bottom panel shows the LD patterns for the same region in a population with stronger and longer LD.

6. FINE-MAPPING

The successful identification of a biological association that has been inspected to be free from statistical artifacts indicates the presence of a functional polymorphism in the neighborhood. The next phase of the experiment is thus to localize this disease-causing variant amongt the polymorphisms in the vicinity, and this often relies on understanding the structure of linkage disequilibrium (LD) between surrounding markers from dense reference databases that are representative of the studied population. The identified polymorphism is very seldom the actual disease-causing variant, but is more likely to be correlated, or in LD, with the functional polymorphism. Given that commercial arrays do not assay every polymorphism in a genomic region, it is not possible to identify all the surrounding genetic variants that are correlated with the identified marker. However, identifying regions of strong LD where the identified marker lies is useful for defining the boundaries of subsequent fine-mapping experiments. This is achieved using LD heatmaps to represent graphically the correlations between all possible pairs of SNPs in a region (Figure 5), often using SNPs from the HapMap, and to investigate how the LD with the identified marker decays with distance. In a population or a genomic region where LD decays more rapidly (upper panel of Figure 5), a smaller region is required for follow-up experiments whereas the presence of longer LD indicates that more SNPs need to be assayed (lower panel of Figure 5) in order to localize the functional variant.

7. CONCLUSION

Never before has the use of well-devised exploratory tools involving numerical summaries and graphical displays been so crucial in aiding the analysis and interpretation of a data set. Traditional data issues like missingness and erroneous entries continue to plague the analysis of data from a GWAS, but the scale of the experiment means that any minor noise or bias in the data introduces unnecessary statistical signals

that complicate the design of follow-up replication or fine-mapping experiments. While false signals that originate from genotyping errors can be identified upon inspecting the clusterplots, associations as a result of cryptic relatedness between the samples, either due to sample handling issues or population structure, cannot be identified from assessing the genotyping but require specialized exploratory tools. A summary of the exploratory diagnostics for a GWAS is given in Table 1, although this is not nor meant to be exhaustive, as more sophisticated tools for EDA are expected to be developed, given the increased understanding of managing and analyzing large-scale genetic data sets.

The next phase of genomic research will aim to progress from identifying associations to establishing the causal mechanisms of diseases. While Section 6 has provided a brief discussion on this process of fine-mapping the functional polymorphisms, the actual implementation is likely to require the integration of phenotype data with deep sequencing data. Sequencing, or the assaying of every individual position in a prescribed genomic region, introduces different challenges in terms of data exploration and analysis. As next generation sequencing technologies assay each genomic region in piecewise fashions, assembling the resultant pieces or "reads" to arrive at the full genomic sequence can require careful quantification and interpretation of the integrity of the information provided by each read. Understanding the methodological challenges in next generation sequencing forms a vital component in the highly anticipated 1000 Genomes Project (http://www.1000genomes.org), which aims to provide high coverage sequence data for about 1200 individuals from around the world. The availability of these genomic sequences is exciting, especially when combined with the process of statistical imputation (Marchini *and others*, 2007), as this suggests the possibility of recovering the sequence data for individuals assayed in genome-wide studies by

Table 1. *Summary of EDA tools for a GWAS*

Type	Analysis	Purpose
Numerical	SNP missingness	To identify and exclude SNPs with high rates of missingness, intended as an indicator of poor genotyping.
Numerical	HWE	To identify and exclude SNPs with extreme deviation from HWE, intended as an indicator of poor genotyping.
Numerical	Minor allele frequency	To identify and exclude SNPs with low minor allele frequencies, intended as an indicator of poor genotyping.
Numerical	Perturbation analysis	Assesses the stability of genotype calls to minor perturbation in hybridization intensities, intended as an indicator for poor genotyping.
Graphical	Clusterplot	Plots the hybridization intensities of all the samples at an SNP with superimposed genotype calls, intended for assessing genotyping accuracy.
Numerical	Sample missingness	To identify and exclude samples with high rates of missingness.
Numerical	Sample heterozygosity	To identify and exclude samples with unusual levels of heterozygosity, intended as an indicator of sample contamination.
Numerical	IBS	To identify pairs of samples with greater extent of concordant genotypes than expected by chance, indicating related or duplicated samples.
Graphical	Principal components biplots	A visual representation of the extent of population structure in the data, useful for defining the type of association analysis required.
Graphical	QQ-plot of test statistic from association analysis	A visual representation of the degree of inflation in the test statistic, indicative of any systematic bias in the data.
Numerical	Genomic control inflation factor	To quantify the degree of inflation of the test statistic, indicative of any systematic bias in the data.
Graphical	LD heatmaps	To represent the correlation between pairs of SNPs in a region, useful for designing replication and fine-mapping experiments.

statistically inferring the unobserved genotypes using the sequence data from the 1000 Genomes Project as reference. This however requires further understanding of the imputation diagnostics to identify genomic regions where imputation confidence is low, which may be attributed to a multitude of reasons including poor coverage of the genome-wide genotyping array, genotyping errors, erroneous sequence assembly, or unexpected genomic differences between the target and reference populations. It is only through careful exploration of the diagnostics from sequencing and imputation that artificial and meaningless association signals will be minimized. Whatever the technological and methodological progress may be in the science of large-scale genetic studies, the traditional act of exploring and understanding the data prior to definitive statistical analysis is unlikely to ever diminish in importance.

ACKNOWLEDGMENTS

The author acknowledges support from the Bill and Melinda Gates Foundation, the Wellcome Trust and the Foundation for the National Institutes of Health through the Grand Challenges in Global Health. *Conflict of Interest:* None declared.

REFERENCES

BALDING, D. J. AND NICHOLS, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identify and paternity. *Genetica* **96**, 3–12.

CARVALHO, B., BENGTSSON, H., SPEED, T. P. AND IRIZARRAY, R. A. (2007). Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data. *Biostatistics* **8**, 485–499.

CLAYTON, D. G., WALKER, N. M., SMYTH, D. J., PASK, R., COOPER, J. D., MAIER, L. M., SMINK, L. J., LAM, A. C., OVINGTON, N. R., STEVENS, H. E. *and others* (2005). Population structure, differential bias and genomic control in a large-scale case-control association study. *Nature Genetics* **37**, 1243–1246.

DEVLIN, B. AND ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55**, 997–1004.

HIRSCHHORN, J. N. AND DALY, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Review Genetics* **6**, 95–108.

INTERNATIONAL HAPMAP CONSORTIUM (2007). A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861.

KORN, J. M., KURUVILLA, F. G., MCCARROLL, S. A., WYSOKER, A., NEMESH, J., CAWLEY, S., HUBBELL, E., VEITCH, J., COLLINS, P. J., DARVISHI, K. *and others* (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genetics* **40**, 1253–1260.

MARCHINI, J., CARDON, L. R., PHILLIPS, M. S. AND DONNELLY, P. (2004). The effects of human population structure on large genetic association studies. *Nature Genetics* **36**, 512–517.

MARCHINI, J., HOWIE, B., MYERS, S., MCVEAN, G. AND DONNELLY, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics* **39**, 906–913.

NCI-NHGRI WORKING GROUP ON REPLICATION IN ASSOCIATION STUDIES (2007). Replicating genotype-phenotype associations. *Nature* **447**, 655–660.

PATTERSON, N., PRICE, A. L. AND REICH, D. (2006). Population structure and eigenanalysis. *PLoS Genetics* **2**, e190.

PLAGNOL, V., COOPER, J. D., TODD, J. A. AND CLAYTON, D. G. (2007). A method to address differential bias in genotyping in large-scale association studies. *PLoS Genetics* **3**, e74.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. AND REICH, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics* **38**, 904–909.

PRITCHARD, J. K., STEPHENS, M., ROSENBERG, N. A. AND DONNELLY, P. (2000). Association mapping in structured populations. *American Journal of Human Genetics* **67**, 170–181.

TEO, Y. Y., FRY, A. E., CLARK, T. G., TAI, E. S. AND SEIELSTAD, M. (2007). On the usage of HWE for identifying genotyping errors. *Annals of Human Genetics* **71**, 701–703.

TEO, Y. Y., INOUYE, M., SMALL, K. S., FRY, A. E., POTTER, S. C., DUNSTAN, S. J., SEIELSTAD, M., BARROSO, I., WAREHAM, N. J., ROCKETT, K. A. *and others* (2008). Whole genome-amplified DNA: insights and imputation. *Nature Methods* **5**, 279–280.

TEO, Y. Y., INOUYE, M., SMALL, K. S., GWILLIAM, R., DELOUKAS, P., KWIATKOWSKI, D. P. AND CLARK, T. G. (2007). A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741–2746.

TEO, Y. Y., SMALL, K. S., CLARK, T. G. AND KWIATKOWSKI, D. P. (2008). Perturbation analysis: a simple method for filtering SNPs with erroneous genotyping in genome-wide association studies. *Annals of Human Genetics* **72**, 368–374.

TUKEY, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

WANG, W. Y., BARRATT, B. J., CLAYTON, D. G. AND TODD, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Review Genetics* **6**, 109–118.

WELLCOME TRUST CASE CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678.

WRIGHT, S. (1943). Isolation by distance. *Genetics* **28**, 114–138.

[*Received January 5, 2009; first revision April 6, 2009; second revision July 22, 2009; accepted for publication September 14, 2009*]