A robust clustering algorithm for identifying problematic samples in genome-wide association studies

Céline Bellenguez[1], Amy Strange[1], Colin Freeman[1], Wellcome Trust Case Control Consortium[†], Peter Donnelly[1,2] and Chris C.A. Spencer[1,*]

[1]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN and
[2]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK

Associate Editor: Jeffrey Barrett

ABSTRACT

Summary: High-throughput genotyping arrays provide an efficient way to survey single nucleotide polymorphisms (SNPs) across the genome in large numbers of individuals. Downstream analysis of the data, for example in genome-wide association studies (GWAS), often involves statistical models of genotype frequencies across individuals. The complexities of the sample collection process and the potential for errors in the experimental assay can lead to biases and artefacts in an individual's inferred genotypes. Rather than attempting to model these complications, it has become a standard practice to remove individuals whose genome-wide data differ from the sample at large. Here we describe a simple, but robust, statistical algorithm to identify samples with atypical summaries of genome-wide variation. Its use as a semi-automated quality control tool is demonstrated using several summary statistics, selected to identify different potential problems, and it is applied to two different genotyping platforms and sample collections.

Availability: The algorithm is written in R and is freely available at www.well.ox.ac.uk/chris-spencer

Contact: chris.spencer@well.ox.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The advent of new technologies, which can simultaneously genotype hundreds of thousands of single nucleotide polymorphisms (SNPs) across the genome, has permitted large-scale studies of human genetic variation. A major application of these technologies is to undertake genome-wide association studies (GWAS) to identify SNPs that correlate with phenotypes such as disease. An important step in providing convincing evidence of association is to argue that the observed correlation is not an artefact of either the sampling strategy (for example, hidden population structure) or systematic biases in inferring genotypes (for example, differences in call rates). In doing so, it has become standard practice to calculate summaries of genome-wide variation that are not expected to vary

systematically between study individuals, and then to identify and remove outlying individuals.

Under the correct statistical model, losing data (that is collected at some expense) nearly always results in reduced statistical power to detect real effects. However, when the model fails to capture the data generating process, inclusion of outlying individuals often leads to an increase in false positives. Exclusion of individuals prior to analysis is a trade-off between loss of power due to reduced sample size, and the benefit of controlling the number of false positives.

The typical approach to identify potentially problematic samples is to calculate summary statistics of genome-wide data and then, by visualizing their distributions across individuals, to manually choose a threshold based on their values for the majority of the data. To automate this process requires an algorithm to infer the distribution of 'normal' study individuals, therefore allowing inference of outliers. For the approach to be applicable in many settings (different summary statistics, genotyping platforms and sample collections) requires a robust model for the outlying individuals.

2 METHODS

Inference of outliers

We implemented a simple mixture model to identify individuals with atypical genome-wide patterns of diversity as measured by m summary statistics of their genotypes or SNP assay intensities data X; $S_1(X_i),\dots,S_m(X_i)$ $(i=1,\dots,n$ with n the number of individuals). To do so, we assume each individual is either 'normal' or an 'outlier', which we index by $Z_i \in \{0,1\}$, and use a Bayesian approach to infer the posterior probability of each individual's membership to the two classes. As summary statistics are averages of many (typically over 500,000) SNPs or assays, the central limit theorem should apply to these statistics across individuals. We consider the distribution of the m summary statistics to be sufficiently well described by independent Gaussian distributions in both the normal and outlier class so that

$$S_j(X_i)|Z_i, \mu_{Z_i,j}, \sigma^2_{Z_i,j} \sim N(\mu_{Z_i,j}, \sigma^2_{Z_i,j}).$$

Having observed the summary statistics, our knowledge of which individuals are outliers is given by the posterior distribution

$$P(Z|S(X)) \propto \int P(S(X)|Z,\mu,\sigma^2) P(Z,\mu,\sigma^2)\, d\mu\, d\sigma^2,$$

where $P(Z,\mu,\sigma^2)$ is the prior distribution. Integrals of this form arise commonly in Bayesian statistics, and it is often not possible to compute them directly. However, there are efficient Monte Carlo methods to sample from the distribution of the unobserved labels Z and the model parameters, μ and σ^2, conditional on the observed data $S(X)$:

$$P(Z,\mu,\sigma^2|S(X)).$$

*To whom correspondence should be addressed.
[†]A list of participants and affiliations appear in Supplementary Material.
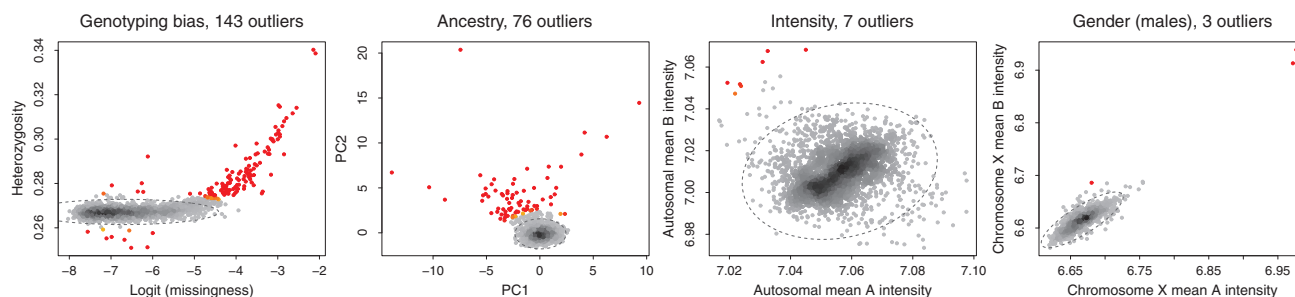
Fig. 1. Outlier identification for 2918 58C samples genotyped on Affymetrix Genome-Wide Human SNP 6.0. 'Normal' individuals are coloured from black to grey, with darker colours denoting higher density of individuals. Outliers are coloured from orange to red, with redder colours denoting higher posterior probability of being an outlier. The 99% confidence ellipse of the inferred distribution of 'normal' individuals is shown as a dashed grey line.

We used Gibbs sampling to obtain T samples from this joint posterior distribution. The posterior probability of the *i*-th individual being an outlier is then estimated as

$$\frac{1}{T}\sum_{t=1}^{T}I(Z_i^{(t)}=1),$$

where $Z_i^{(t)}$ is the class membership of the *i*-th individual for the sample *t* and *I* is the indicator function. An individual is then considered as an outlier if its estimated posterior probability of being an outlier is >50%.

This approach easily generalizes to correlated summary statistics. Here we consider only two summary statistics jointly, but the model could be extended to more. Information on either the distribution of summary statistics for normal individuals, perhaps from previous analysis, or the fraction of individuals which are outlying can both be specified through prior distributions. See Supplementary Material for details.

Approximation for robust detection of outliers

To facilitate the use of Gibbs sampling, we used conjugate priors for the model parameters, except for the variance of the outlier class. To ensure identifiability, we assume that the SD of the outliers (for which $Z_i = 1$) is factor λ larger than the SD of the normal individuals (for which $Z_i = 0$) so that:

$$\sigma_{1,j}^2 = \lambda^2 \sigma_{0,j}^2.$$

The parameter λ is fixed *a priori* and controls the stringency of the outlier classification. Using this hard prior assumption, the variance of outlier class is completely determined by the variance of the normal class. We made the additional assumption that the percentage of outlier samples is small, so that all the information about the variance of the normal class is assumed to come from the normal individuals:

$$P(\sigma_{0,j}^2|S_j(X),Z,\mu_{0,j},\mu_{1,j})\approx P(\sigma_{0,j}^2|S_j(X_{Z=0}),Z,\mu_{0,j}).$$

This assumption adds robustness to the model: the distribution of the outliers will have little impact on the model fit which, because of the light tails of the Guassian distribution, can be heavily influenced by outlying observations. The approximation is similar in motivation to the concept of trimmed-likelihoods, where the likelihood is computed after trimming the least likely observations (Hadi and Luceno, 1997) or perhaps also to contamination models, where the influence of the outliers goes to zero.

3 APPLICATION

We applied the clustering approach independently to four different control datasets genotyped as part of the Wellcome Trust Case Control Consortium 2 (WTCCC2). These comprised 2918 samples from the 1958 Birth Cohort (58C) and 2530 National Blood Service controls (UKBS) genotyped on the Affymetrix Genome-Wide Human SNP 6.0 and the Illumina custom Human 1.2M-Duo chips. We considered four different quality control criteria, based on summaries of each individual's genotypes or probe intensities:

- *Genotyping bias*: genome-wide heterozygosity (the fraction of heterozygote calls) and call rate (the proportion of missing genotypes). Indicative of assay failure, contamination or inbreeeding.
- *Ancestry*: projection of individual's genotypes onto two axes of variation which differentiate individuals with European, Asian and African ancestry. Indicative of individuals with atypical ancestry with respect to the majority of the sample.
- *Intensity*: genome-wide average of the probe intensities which target the two alleles at each autosomal SNP. Indicative of partial assay failure or insufficient normalization.
- *Gender*: for females and males separately, the mean probe intensities across SNPs on chromosome X. Indicative of incorrect gender assignment.

Results are shown in Figure 1 and Supplementary Figures S1–S3 for, 58C and UKBS samples genotyped on Affymetrix and Illumina platforms, respectively. As well as being statistically principled, in practice, it is helpful that, once the prior distributions have been specified, identification of outliers is automatic. Empirically, it appears to make sensible inference for a range of normal and outlier distributions, suggesting it is useful for quality control in GWAS [successfully applied in, for example Genetic Analysis of Psoriasis Consortium & the WTCCC2 (2010); The International Multiple Sclerosis Genetics Consortium & the WTCCC2 (2011); The UK IBD Genetics Consortium & the WTCCC2 (2009)] and perhaps in other settings.

Funding: This work was supported by Wellcome Trust awards 090532/Z/09/Z, 075491/Z/04/B and 084575/Z/08/Z. PD was supported in part by a Royal Society Wolfson Merit Award and CCAS by a Nuffield Department of Medicine Scientific Leadership Fellowship.

Conflict of Interest: none declared.

REFERENCES

Hadi,A.S. and Luceno,A (1997) Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Comput. Stat. Data Anal.*, **25**, 251–272.

Genetic Analysis of Psoriasis Consortium & the WTCCC2 (2010) A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.*, **42**, 985–990.

The International Multiple Sclerosis Genetics Consortium & the WTCCC2 (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*, **476**, 214–219.

The UK IBD Genetics Consortium & the WTCCC2 (2009) Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.*, **41**, 1330–1334.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016