Data quality control in genetic case-control association studies

Carl A Anderson[1,2], Fredrik H Pettersson[1], Geraldine M Clarke[1], Lon R Cardon[3], Andrew P Morris[1] & Krina T Zondervan[1]

[1]Genetic and Genomic Epidemiology Unit, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK. [2]Statistical Genetics, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK. [3]GlaxoSmithKline, King of Prussia, Pennsylvania, USA. Correspondence should be addressed to C.A.A. (carl.anderson@sanger.ac.uk) or K.T.Z. (krinaz@well.ox.ac.uk).

Published online 26 August 2010; doi:10.1038/nprot.2010.116

This protocol details the steps for data quality assessment and control that are typically carried out during case-control association studies. The steps described involve the identification and removal of DNA samples and markers that introduce bias. These critical steps are paramount to the success of a case-control study and are necessary before statistically testing for association. We describe how to use PLINK, a tool for handling SNP data, to perform assessments of failure rate per individual and per SNP and to assess the degree of relatedness between individuals. We also detail other quality-control procedures, including the use of SMARTPCA software for the identification of ancestral outliers. These platforms were selected because they are user-friendly, widely used and computationally efficient. Steps needed to detect and establish a disease association using case-control data are not discussed here. Issues concerning study design and marker selection in case-control studies have been discussed in our earlier protocols. This protocol, which is routinely used in our labs, should take approximately 8 h to complete.

INTRODUCTION

Shortcomings in study design and errors in genotype calling have the potential to introduce systematic biases into genetic case-control association studies, leading to an increase in the number of false-positive and false-negative associations (see **Box 1** for a glossary of terms). Many such errors can be avoided through a careful selection of case and control groups and vigilant laboratory practices. A protocol for the successful ascertainment of unbiased case-control groups, focusing on sampling individuals from the same underlying population, is provided in an earlier publication in this series[1]. In the current protocol, we assume that these guidelines regarding sample ascertainment have been followed. However, even when case-control association study design has been conducted appropriately, a thorough assessment of data quality—including testing whether 'same-population sampling' was successful—should still be undertaken. Such assessments allow the identification of substandard markers and samples, which should be removed before further analysis to reduce the number of false-positive and false-negative associations.

This protocol deals with the quality control (QC) of genotype data from genome-wide and candidate-gene case-control association studies, and outlines the methods routinely used in key studies from our research groups[2,3]. Although the protocol applies to genotypes after they have been determined ('called') from probe intensity data, it is still important to understand how the genotype calling was conducted. Traditionally, for small-scale genotyping efforts, manual inspection of allele-probe intensities is carried out to call genotypes, and this is still the situation for many candidate-gene or replication studies conducted at present. However, when undertaking genome-wide association (GWA) studies with so many markers, this is no longer practical. Genotype-calling algorithms (implemented in proprietary software accompanying the genotyping platform, or used externally in software such as Illuminus[4] or Chiamo[2]) use mathematical clustering algorithms to analyze raw intensity data and—for a given individual and a given marker locus—to estimate the probability that the genotype is *aa*, *Aa* or *AA*. A threshold is then applied such that any genotype with a probability exceeding a certain cutoff is accepted and referred to as 'called'; otherwise the genotype is not accepted and is referred to as 'uncalled' or 'missing'. The threshold applied can substantially affect the genotype call rate and the quality of the genotype data. If it is set too low and the separation of signal clouds is poor, erroneous genotypes can be assigned. However, calling only genotypes with high certainty can result in 'informative missingness' because failure to call may be dependent on genotype. For example, rare homozygous genotypes may, on average, have lower probabilities, thus introducing bias to allele frequencies based only on called genotypes[5]. Furthermore, a high calling threshold will produce a large proportion of unnecessarily uncalled (missing) genotypes, thus reducing genomic coverage and the power to detect association. The ultimate assessment of genotype quality is manual inspection of cluster plots, and it is essential that after association testing these plots are inspected for all SNPs taken forward for replication regardless of QC intensity (to prevent wasteful replication efforts).

Genome-wide association

Because of the large number of marker loci tested for association in a GWA study, even a low rate of error or bias can be detrimental. If 1 million markers are tested for association and the proportion of poorly genotyped markers is 0.001, then—if the inaccurate calling results in the detection of a spurious association—up to 1,000 markers may be unnecessarily taken forward for replication because of a false-positive association. In an attempt to remove these false-positive associations, one must undertake several QC steps to remove individuals or markers with particularly high error rates. If, as advised, many thousands of cases and controls have been genotyped to maximize the power to detect association, the removal of a handful of individuals should have little effect on overall power. Furthermore, given the large number of markers genotyped in modern GWA studies, the removal of a (hopefully) small percentage of these should not markedly decrease

BOX 1 | GLOSSARY

Cochran-Armitage trend test—Statistical test for analysis of categorical data when categories are ordered. It is used to test for association in a 2 × *k* contingency table (where *k* is the number of exposure categories—in this case, genotype classes). In genetic association studies, because the underlying genetic model is unknown, the additive version of this test is most commonly used.
Confounding—A type of bias in statistical analysis causing spurious or distorted findings caused by a correlation between an extraneous variable (the confounding variable) and both the exposure variable (e.g., the genotype at a given locus) and the outcome variable (e.g., case-control status).
Failure rate—The proportion of missing genotypes. Genotypes are classified as missing if the genotype-calling algorithm cannot infer the genotype with sufficient confidence. Can be calculated across individuals and/or SNPs.
False negative—This occurs when a true disease-associated variant is not associated with disease in a given study.
False positive—This occurs when a variant not genuinely associated with disease status is significantly associated with disease in a given study.
Genotype-calling algorithm—A statistical algorithm that, per marker and per individual, converts intensity data from two allelic probes into a single genotype for analysis.
Genotype call rate—The proportion of genotypes per marker with nonmissing data.
Genotype cluster plots—Per-SNP graphical representations of intensity data from two probes used during genotyping across all individuals, together with the final called genotype. Typically, across all individuals, the intensity of probe A is plotted against the intensity of probe B and the genotype for a given individual is represented by one of three colors. Genotypes of the same class should cluster together and these clusters should be consistent across case and control groups.
HapMap—An international project to create a haplotype map of the human genome. The publicly available data consist of ~3.2 million SNPs genotyped across four sample sets of 60–90 individuals of African, Asian or European ancestry (stage II). HapMap Phase III consists of ~1.5 million SNP genotypes from a greater number of individuals and populations.
Hardy-Weinberg equilibrium—Given a minor allele frequency of q, the probabilities of the three possible genotypes (*aa*, *Aa* and *AA*) at a biallelic locus that is in Hardy-Weinberg equilibrium are $((1-q)^2, 2q(1-q), q^2)$. In a large, randomly mating, homogenous population, these probabilities should be stable from generation to generation.
Heterozygosity rate—The proportion of heterozygous genotypes for a given individual.
Informative missingness—This occurs when the probability of a genotype being called missing is correlated with the true underlying genotype.
Linkage disequilibrium—Nonrandom association of alleles at two or more loci.
Pairwise identity by state—The proportion of loci in which a given pair of individuals share the same alleles. Given by (IBS2 + 0.5 × IBS1)/(*N* SNP pairs), where IBS2 and IBS1 are the number of loci in which the two individuals have two alleles and one allele in common, respectively, and *N* SNP pairs is the number of common, nonmissing SNPs.
Population substructure—The presence of distinct groups of individuals with subtle differences in allele frequency such that genetic data can be used to cluster these individuals into separate groups.
Principal component analysis—A mathematical procedure for calculating a number of orthogonal latent variables that summarize a data matrix containing many potentially correlated variables.
r^2—A measure of the linkage disequilibrium (genetic correlation) between two markers. An r^2 value of 1 indicates that the two markers are perfectly correlated and an r^2 value of 0 indicates that the two markers are completely independent.

the overall power of the study. That said, every marker removed from a study is potentially an overlooked disease association and thus the impact of removing one marker is potentially greater than the removal of one individual (although genotype imputation can be used to recover these markers[6]). In this protocol, we advocate implementing QC on a 'per-individual' basis before conducting QC on a 'per-marker' basis to maximize the number of markers remaining in the study. This approach prevents markers from being erroneously removed because of a subset of poorly genotyped individuals, but individuals may be falsely removed on the basis of a poorly genotyped subset of markers. An alternative (and conservative) approach would be to complete both QC stages before removing any individuals or markers, but data may then be removed unnecessarily.

Per-individual QC

Per-individual QC of GWA data consists of at least four steps: (i) identification of individuals with discordant sex information, (ii) identification of individuals with outlying missing genotype or heterozygosity rates, (iii) identification of duplicated or related individuals and (iv) identification of individuals of divergent ancestry.

It is advantageous to begin by using genotype data from the X chromosome to check for discordance with ascertained sex and thus highlight plating errors and sample mix-ups. Because males have only one copy of the X chromosome, they cannot be heterozygous for any marker that is not in the pseudoautosomal region of the Y chromosome. Typically, when a genotype-calling algorithm detects a male heterozygote for an X-chromosome marker, it calls that genotype as missing. Therefore, female DNA samples that are marked as male in the input files will have a large amount of missing data because all their heterozygous X-chromosome genotypes will be set to missing. Not all genotype-calling algorithms automatically set heterozygous haploid genotypes to missing; by calling X-chromosome markers 'blind' to ascertained sex, this functionality can be removed in those that do. Typically, one expects male samples to have a homozygosity rate of 1 (although, because of genotyping error, this can vary) and females to have a homozygosity rate of <0.2. Male DNA samples that are marked

as female in input files will have a higher-than-expected homozygosity rate and female samples marked as male will have a lower-than-expected heterozygosity rate. Therefore, the best way to detect discrepancies between genotype information and ascertained sex is to calculate homozygosity rates across all X-chromosome SNPs for each individual in the sample and compare these with the expected rates. The sex of a case or control is typically only of relevance when these data are to be used during analysis; for example, when carrying out a sex-stratified analysis or when analyzing the X chromosome. However, when a sample with discordant sex information is detected, it is important that it is investigated to ensure that another DNA sample has not been genotyped by mistake (because the wrong (sub)phenotype data may be connected to the genotypes). Unless the sample can be correctly identified using existing genotype data or it can be confirmed that sex was recorded incorrectly, individuals with discordant sex information should be removed from further analysis.

Large variations exist in DNA sample quality and these can have substantial effects on genotype call rate and genotype accuracy. Samples of low DNA quality or concentration often have below-average call rates and genotype accuracy. The genotype failure rate and heterozygosity rate per individual are both measures of DNA sample quality. Typically, individuals with more than 3–7% missing genotypes are removed[2,7]. Carefully scrutinizing the distribution of missing genotype rates across the entire sample set is the best way to ascertain the most appropriate threshold. Similarly, the distribution of mean heterozygosity (excluding sex chromosomes) across all individuals should be inspected to identify individuals with an excessive or reduced proportion of heterozygote genotypes, which may be indicative of DNA sample contamination or inbreeding, respectively. Mean heterozygosity (which is given by $(N-O)/N$, where N is the number of non-missing genotypes and O the observed number of homozygous genotypes for a given individual) will differ between populations and SNP genotyping panels. Owing to the increased success rate and accuracy of modern high-throughput genotyping methodologies (including genotype-calling algorithms), typically, these measures jointly lead to only a small proportion of individuals being excluded from further analysis.

A basic feature of standard population-based case-control association studies is that all samples are unrelated (i.e., the maximum relatedness between any pair of individuals is less than that of a second-degree relative). If duplicates, first- or second-degree relatives are present, a bias may be introduced in the study because the genotypes within families will be overrepresented, and thus the sample may no longer be a fair reflection of the allele frequencies in the entire population. In population-based case-control studies, all efforts should be made to limit the number of duplicate and related individuals in the design phase of the study (although the deliberate inclusion of duplicate samples can be used to determine the genotyping error rate)[8]. To identify duplicate and related individuals, calculate a metric (identity by state, IBS) for each pair of individuals based on the average proportion of alleles shared in common at genotyped SNPs (excluding sex chromosomes). The method works best when only independent SNPs are included in the analysis. To achieve this, regions of extended linkage disequilibrium (LD) (such as the human leukocyte antigen (HLA) region) are removed entirely from the data set[9] and remaining regions are typically pruned so that no pair of SNPs within a given window (e.g., 50 kb) is correlated (typically taken as $r^2 > 0.2$). Following the calculation of IBS between all pairs of individuals, duplicates are denoted as those with an IBS of 1. The population's mean IBS will vary depending on the allele frequency of genotyped markers within that population. Related individuals will share more alleles IBS than expected by chance, with the degree of additional sharing proportional to the degree of relatedness. The degree of recent shared ancestry for a pair of individuals (identity by descent, IBD) can be estimated with genome-wide IBS data using software such as PLINK[10]. The expectation is that IBD = 1 for duplicates or monozygotic twins, IBD = 0.5 for first-degree relatives, IBD = 0.25 for second-degree relatives and IBD = 0.125 for third-degree relatives. Owing to genotyping error, LD and population structure, there is often some variation around these theoretical values and it is typical to remove one individual from each pair with an IBD value of > 0.1875, which is halfway between the expected IBD for third- and second-degree relatives. For these reasons an IBD value of > 0.98 identifies duplicates.

Confounding can be a major source of bias in population-based case-control studies and is caused by underlying differences between the case and control subgroups other than those directly under study (typically, disease status) that correlate with the exposure variable. In the case of genetic studies, in which the exposure of interest is genotype distribution, a major source of confounding is population stratification, in which genotypic differences between cases and controls are generated because of different population origins rather than because of any effect on disease risk[11]. For example, Campbell *et al.*[12] carried out an association analysis on a panel of European-American individuals discordant for height and detected significant association with *LCT*, a locus that has undergone strong selection in certain European populations; the frequency of variants within this gene differs markedly between populations. After cases and controls were matched for population ancestry, the evidence of association at this locus decreased significantly. Although a well-designed population case-control study attempts to draw cases and controls from the same population, a hidden fine-scale genetic substructure within that single population (or the inadvertent inclusion of individuals from another population) cannot be ruled out. Confounding occurs when the population substructure is not equally distributed between case and control groups. In this scenario, a signal of association will arise for an ancestrally informative SNP, not because of an association with disease risk but because of allele frequency differences between the founder populations that differentially comprise cases and controls. Even a small degree of population stratification can adversely affect a GWA study because of the large sample sizes required to detect common variants underlying most complex diseases[13]. Therefore, after giving careful consideration to matching of cases and controls on population origin[1], potential stratification must be examined and characterized during QC. Efforts should then be made to remove or reduce the effect of population stratification through the removal of individuals of divergent ancestry. Correction for fine-scale or within-population substructure can be attempted during association testing, but this is beyond the scope of this protocol.

The most common method for identifying (and subsequently removing) individuals with large-scale differences in ancestry is principal component analysis (PCA)[14,15]. An alternative yet related method, multidimensional scaling (implemented in PLINK), is

available but requires the construction of a pairwise IBD matrix and is therefore more computationally complex. PCA is a multivariate statistical method used to produce several uncorrelated variables (or principal components) from a data matrix containing observations across a number of potentially correlated variables. The principal components are calculated so that the first principal component accounts for as much variation as possible in the data in a single component; this is followed by the second component and so on. In the PCA model of ancestry detection, the observations are the individuals and the potentially correlated variables are the markers. A principal component model is built using pruned genome-wide genotype data from populations of known ancestry; for example, to detect large-scale (continental-level) ancestry, one could use the HapMap genotype data from Europe (CEU), Asia (CHB + JPT) and Africa (YRI)[16,17]. Because of the large-scale genetic differences between these three ancestral groups, the first two principal components are sufficient to separately cluster individuals from the three populations. The PCA model can then be applied to the GWA individuals to predict principal component scores for these samples, thus allowing them to be clustered in terms of ancestry alongside the HapMap samples. A common set of approximately 50,000 independent markers must be used for the model-building and prediction steps. Regions of extended high LD (such the HLA region) should be removed before the analysis because these can overly influence the principal component model[9]. The PCA method can also be used to cluster individuals based on fine-scale structure, although more principal components may be needed to capture this variation fully, and appropriate reference samples will be required.

Per-marker QC

Per-marker QC of GWA data consists of at least four steps: (i) identification of SNPs with an excessive missing genotype, (ii) identification of SNPs showing a significant deviation from Hardy-Weinberg equilibrium (HWE), (iii) identification of SNPs with significantly different missing genotype rates between cases and controls and (iv) the removal of all markers with a very low minor allele frequency (MAF).

The removal of suboptimal markers is key to the success of a GWA study, because they can present as false positives and reduce the ability to identify true associations correlated with disease risk. However, the criteria used to filter out low-quality markers differ from study to study. Great care must be taken to remove only poorly characterized markers because every removed marker is potentially a missed disease variant. Classically, markers with a call rate less than 95% are removed from further study[7,18], although some studies have chosen higher call-rate thresholds (99%) for markers of low frequency (MAF $<5\%$)[2].

Most GWA studies exclude markers that show extensive deviation from HWE because this can be indicative of a genotyping or genotype-calling error. However, deviations from HWE may also indicate selection; accordingly, if a case sample shows deviations from HWE at loci associated with disease, it would obviously be counterproductive to remove these loci from further investigation[19]. Therefore, only control samples should be used when testing for deviations for HWE. The significance threshold for declaring SNPs to be in Hardy-Weinberg equilibrium has varied significantly between studies (P-value thresholds between 0.001 and 5.7×10^{-7} have been reported in the literature[2,20]). However, those studies that have set very low thresholds for HWE deviations have done so on the condition that all genotype cluster plots for SNPs showing some evidence of deviation from HWE (i.e., $P<0.001$) are examined manually for quality. In practice, this means that many SNPs with an HWE P-value less than 0.001 are removed, although robustly genotyped SNPs below this threshold remain under study.

Testing for, and subsequently removing, SNPs with substantial differences in missing genotype rates between cases and controls is another means of reducing confounding and removing poorly genotyped SNPs[21]. Calling case and control genotypes together, or using 'fuzzy calls'[22], greatly reduces this confounding, but important differences in genotype failure may still exist in the data and present as false-positive associations. In studies in which cases and/or controls have been drawn from several different sources, it is wise to test for major differences in call rate, allele frequency and genotype frequency between these various groups to ensure that it is fair to treat the combined case or control set as one homogenous group.

The final step in per-marker QC is to remove all SNPs with a very low MAF. Typically, a MAF threshold of 1–2% is applied, but studies with small sample sizes may require a higher threshold. The small sizes of the heterozygote and rare-homozygote clusters make these variants difficult to call using current genotype-calling algorithms; they frequently present as false positives in case-control association tests. Furthermore, even when accurately called, association signals observed at these rare SNPs are less robust because they are driven by the genotypes of only a few individuals. Given that the power to detect association with rare variants is so low[23], their removal does not overly affect the study. However, even after the removal of rare variants and stringent individual and SNP QC, genotyping errors may still persist. Checking cluster plots manually is the best way to ensure that genotype calls are robust. Therefore, it is essential that all SNPs associated with disease status be manually inspected for clustering errors before choosing SNPs for follow-up genotyping.

Candidate-gene association

Candidate-gene association studies involve far fewer SNPs than GWA studies, and therefore many of the GWA study QC procedures cannot be undertaken in candidate-gene studies. One of the advantages of the GWA study approach is that more than 99% of SNPs follow the null distribution of no association and can be used to detect evidence of confounding. This is not possible in a candidate-gene approach because (i) owing to the gene's candidacy there may be few SNPs falling under the null hypothesis of no association and (ii) far fewer SNPs are genotyped. With fewer genotyped SNPs, it is also more difficult to obtain accurate estimates of (i) DNA quality (through genotype failure rate and heterozygosity rate), (ii) population ancestry and (iii) familial relationships with others in the study. The detrimental effect these factors can have on a candidate-gene association may be equal to that observed in a GWA scenario (although in a well-designed study of ethnically matched individuals, the prior probability of population stratification at a single locus is much lower than that of stratification at any locus across the genome) except that the researcher's ability to identify and remove erroneous individuals and SNPs is greatly reduced. This is perhaps another reason why candidate-gene studies have typically not yielded many reproducible disease gene associations (in addition to small sample sizes, poor coverage of genetic variation and poor choice of candidates).

One should still attempt to identify and remove individuals with exceptionally low call rates. However, the threshold at which individuals are excluded varies depending on the number of SNPs genotyped and is typically higher than that used when carrying out a GWA study. For example, if a candidate-gene study includes 50 SNPs, then removing individuals with more than 3% missing data would result in removing individuals missing as few as 2 SNPs. A more reasonable approach would be to remove those individuals missing 10 or more SNPs (a failure rate of 0.2).

The QC of markers in candidate-gene studies is more comparable to that for the GWA study approach, as similar numbers of cases and controls are involved. It is extremely important to examine the failure rate of markers included in the candidate-gene study and exclude those with a high failure rate. When a SNP is identified with a high failure rate (> 5%), an option is to return to the laboratory and attempt to regenotype that SNP in individuals with missing data. Given that SNPs included in a candidate-gene study are chosen on the basis of their ability to tag neighboring SNPs[24], the exclusion of a SNP because of an elevated failure rate can seriously impair a candidate-gene study. With this in mind, when a SNP is not genotyped with sufficient quality across individuals (and this cannot be rectified by regenotyping), it is advisable to return to the design stage and select another tag for the haplotype block in which the failed SNP resides. Detection of deviations from HWE in controls is still a relevant method for checking genotyping quality.

Software

Standard statistical software (such as R[25] or SPSS) can be used to conduct all the analyses outlined above. However, for GWA studies, many researchers choose to use custom-built, freely available software such as PLINK[10], GenABEL[26], GS2[27] or snpMatrix (an R package that forms part of the bioconductor project, http://www.bioconductor.org/). These software programs can also be used for candidate-gene association studies. The advantages of these compared with standard statistical software are that (i) they store the large genome-wide SNP data in memory-efficient data structures, thus significantly improving computational efficiency and reducing disk usage and (ii) they more fully automate many of the necessary analyses. Although PLINK is used in the present protocol, all of the other packages are equally suitable.

The next section describes protocols for QC of GWA and candidate-gene data. The protocol is illustrated using simulated data sets, which are available for download at http://www.well.ox.ac.uk/~carl/gwa/nature-protocols (see also **Supplementary Data**).

MATERIALS

EQUIPMENT

Data

- Genome-wide (raw-GWA-data.tgz) and candidate-gene (raw-PPARG-data.tgz) SNP data and software scripts (http://www.well.ox.ac.uk/~carl/gwa/nature-protocols). See also **Supplementary Data** for raw-GWA-data.tgz

Software

- Computer workstation with Unix or Linux operating system
- PLINK software[10] (http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml)
- SMARTPCA.pl[15] software for running PCA (http://genepath.med.harvard.edu/~reich/Software.htm)
- Statistical software for data analysis and graphing, such as R (http://cran.r-project.org/)

PROCEDURE

Creation of BED files ● TIMING ~20 min

1| The format in which genotype data are returned to investigators varies among genome-wide SNP platforms and genotyping centers. We assume that genotypes have been called by a genotyping center and returned in the standard PED and MAP file formats (see **Box 2**). For candidate-gene studies, please see **Box 3**.

2| Type '`tar xfvz raw-GWA-data.tgz`' at the shell prompt to unpack the gzipped .tar file and create the files raw-GWA-data.map and raw-GWA-data.ped.

3| Use PLINK to create BED, BIM and FAM files. Type '`plink --file raw-GWA-data --make-bed --out raw-GWA-data`' at the shell prompt.

▲ **CRITICAL STEP** BED files save data in a more memory- and time-efficient manner (binary files) to facilitate the analysis of large-scale data sets[10]. PLINK creates a .log file (named raw-GWA-data.log) that details (among other information) the implemented commands, the number of cases and controls in the input files, any excluded data and the genotyping rate in the remaining data. This file is very useful for checking whether the software is successfully completing commands.

BOX 2 | PED AND MAP FILES

A PED file is a white space (space or tab)-delimited file in which each line represents one individual and the first six columns are mandatory and in the order 'Family ID', 'Individual ID', 'Paternal ID', 'Maternal ID', 'Sex (1=male, 2=female, 0=missing)' and 'Phenotype (1=unaffected, 2=affected, 0=missing)'. The subsequent columns denote genotypes that can be any character (e.g., 1, 2, 3, 4 or A, C, G, T). Zero denotes a missing genotype. Each SNP must have two alleles (i.e., both alleles are either present or absent). The order of SNPs in the PED file is given in the MAP file, in which each line denotes a single marker and the four white-space–separated columns are chromosome (1–22, X, Y or 0 for unplaced), marker name (typically an rs number), genetic distance in Morgans (this can be fixed to 0) and base-pair position (bp units).

BOX 3 | CANDIDATE-GENE STUDY

Creation of PED and MAP files ● TIMING ~ 5 min

1. The format in which genotype data are returned to investigators depends on where the genotyping was conducted and on which platform. We assume that genotypes have been called by a genotyping center and returned in the standard PED and MAP file formats (see **Box 1**). Download the file 'raw-PPARG-data.tgz'.

2. Type 'tar xfvz raw-PPARG-data.tgz' at the shell prompt to unpack the gzipped .tar file and create the files 'raw-PPARG-data.map' and 'raw-PPARG-data.ped'.

Test markers for different genotype call rates between cases and controls ● TIMING ~ 5 min

3. At the Unix prompt, type 'plink --bfile raw-PPARG-data --test-missing --out raw-PPARG-data' to test all markers for differences in call rates between cases and controls. The output of this test can be found in raw-PPARG-data.missing.

4. At the Unix prompt, type 'perl run-diffmiss-qc.pl clean-inds-GWA-data' to create a file named 'fail-diffmiss-qc.txt', which contains all SNPs with a significantly different (P<0.00001) missing data rate between cases and controls.

Run additional QC steps and remove failing markers and samples ● TIMING ~ 5 min

5. At the Unix prompt, type 'plink --file raw-PPARG-data --exclude fail-diffmiss-qc.txt –mind 0.1 --maf 0.01 --geno 0.05 --hwe 0.00001 --recode --out clean-PPARG-data'. In addition to markers failing previous QC steps, those with a MAF < 0.01 and an HWE P-value < 0.00001 (in controls) are also removed. Samples with a genotype failure rate greater than 0.1 are also removed.

▲ **CRITICAL STEP** If genotypes are not available in PED file format, both GS2[27,28] and PLINK have functionality to read several other file formats and convert these into PED files (or even directly into BED files).

Identification of individuals with discordant sex information ● TIMING ~30 min

4| At the Unix prompt, type 'plink --bfile raw-GWA-data --check-sex --out raw-GWA-data' to calculate the mean homozygosity rate across X-chromosome markers for each individual in the study.

5| Produce a list of individuals with discordant sex data by typing 'grep PROBLEM raw-GWA-data.sexcheck > raw-GWA-data.sexprobs' and open the file to obtain the family IDs (column 1) and individual ID (column 2) for these individuals. Column 3 denotes ascertained sex and column 4 denotes sex according to genotype data. When the homozygosity rate is more than 0.2 but less than 0.8, the genotype data are inconclusive regarding the sex of an individual and these are marked in column 4 with a 0.

6| Report the IDs of individuals with discordant sex information to those who conducted sex phenotyping. In situations in which discrepancy cannot be resolved, add the family ID (FID) and individual ID (IID) of the samples to a file named 'fail-sexcheck-qc.txt' (one individual per line, tab delimited).

Identification of individuals with elevated missing data rates or outlying heterozygosity rate ● TIMING ~30 min

7| At the shell prompt, type 'plink --bfile raw-GWA-data --missing --out raw-GWA-data' to create the files raw-GWA-data.imiss and raw-GWA-data.lmiss. The fourth column in the .imiss file (N_MISS) denotes the number of missing SNPs and the sixth column (F_MISS) denotes the proportion of missing SNPs per individual.

8| At the shell prompt, type 'plink --bfile raw-GWA-data --het --out raw-GWA-data' to create the file raw-GWA-data.het, in which the third column denotes the observed number of homozygous genotypes [O(Hom)] and the fifth column denotes the number of nonmissing genotypes [N(NM)] per individual.

9| Calculate the observed heterozygosity rate per individual using the formula (N(NM) – O(Hom))/N(NM). Create a graph in which the observed heterozygosity rate per individual is plotted on the x axis and the proportion of missing SNPs per individuals is plotted on the y axis. This can be carried out using standard software such as Excel or statistical packages such as SPSS. A script for calculating the heterozygosity rate and producing the graph using R is supplied (imiss-vs-het.Rscript). Type 'R CMD BATCH imiss-vs-het.Rscript' at the Unix prompt to run this script and create the graph (raw-GWA-data.imiss-vs-het.pdf).

10| Examine the plot to decide reasonable thresholds at which to exclude individuals based on elevated missing or extreme heterozygosity. We chose to exclude all individuals with a genotype failure rate ≥0.03 (**Fig. 1**, vertical dashed line) and/or a heterozygosity rate ± 3 s.d. from the mean (**Fig. 1**, horizontal dashed lines). Add the FID and IID of the samples failing this QC to the file named 'fail-imisshet-qc.txt'.

Identification of duplicated or related individuals ● TIMING ~4 h

11| To minimize computational complexity, reduce the number of SNPs used to create the IBS matrix by pruning the data set so that no pair of SNPs (within a given number of base pairs) has an r^2 value greater than a given threshold

(typically, 0.2). Given that our current data set was simulated ignoring LD, this step was not applicable, but a list of SNPs for inclusion in this step (raw-GWA-data.prune.in) can be downloaded from http://www.well.ox.ac.uk/~carl/gwa/nature-protocols.

▲ **CRITICAL STEP** In real data sets in which LD is present, data can be pruned by typing at the shell prompt '`plink --file raw-GWA-data --exclude high-LD-regions.txt --range --indep-pairwise 50 5 0.2 --out raw-GWA-data`' to create the file raw-GWA-data.prune.in; this saves the list of SNPs to be kept in the analysis. This also excludes SNPs from extended regions of high LD listed in `high-LD-regions.txt`.

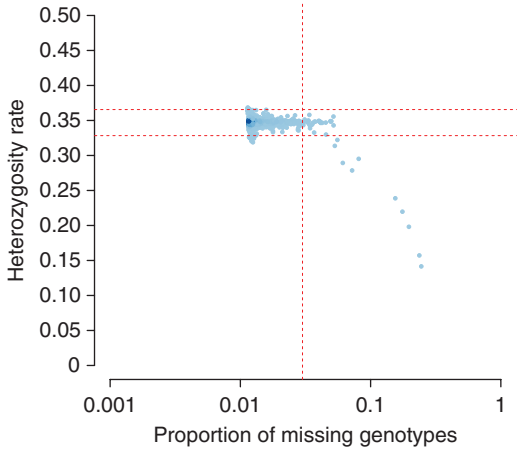

Figure 1 | Genotype failure rate versus heterozygosity across all individuals in the study. Shading indicates sample density and dashed lines denote quality control thresholds.

12| Type '`plink --bfile raw-GWA-data --extract raw-GWA-data.prune.in --genome --out raw-GWA-data`' at the shell prompt to generate pairwise IBS for all pairs of individuals in the study based on the reduced marker set.

▲ **CRITICAL STEP** Because this step can take much time, it is advisable to prefix the above command with the Unix command *nohup* to allow the command to continue running on the machine after the user logs out. Placing an ampersand (&) at the end of the command will free the Unix terminal for further use.

13| Type '`perl run-IBD-QC.pl raw-GWA-data`' at the Unix prompt to identify all pairs of individuals with an IBD > 0.185. The code looks at the individual call rates stored in raw-GWA-data.imiss and outputs the IDs of the individual with the lowest call rate to '`fail-IBD-QC.txt`' for subsequent removal.

Identification of individuals of divergent ancestry ● **TIMING ~2 h**

14| This step is conducted by merging study genotypes to HapMap Phase III (HapMap3) data from four ethnic populations. The alleles at each marker must be aligned to the same DNA strand to allow the study data to merge correctly. Because not all SNPs are required for this analysis, A→T and C→G SNPs, which are more difficult to align, can be omitted. To create a new BED file, excluding from the GWA data those SNPs that do not feature in the genotype data of the four original HapMap3 populations, type '`plink --bfile raw-GWA-data --extract hapmap3r2_CEU.CHB.JPT.YRI.no-at-cg-snps.txt --make-bed --out raw-GWA-data.hapmap-snps`' at the Unix prompt.

15| Merge the `raw-GWA-data.hapmap-snps` files with the HapMap data and extract the pruned SNP set by typing '`plink --bfile raw-GWA-data.hapmap-snps --bmerge hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.bed hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.bim hapmap3r2_CEU.CHB.JPT.YRI.founders.no-at-cg-snps.fam --extract raw-GWA-data.prune.in --make-bed --out raw-GWA-data.hapmap3r2.pruned`'.

? TROUBLESHOOTING

16| Create a copy of BIM and FAM files by typing '`cp raw-GWA-data.hapmap3r2.pruned.bim raw-GWA-data.hapmap3r2.pruned.pedsnp`', followed by '`cp raw-GWA-data.hapmap3r2.pruned.fam raw-GWA-data.hapmap3r2.pruned.pedind`' at the Unix prompt.

17| Conduct a PCA on the merged data by typing '`perl smartpca.perl -i raw-GWA-data.hapmap3r2.pruned.bed -a raw-GWA-data.hapmap3r2.pruned.pedsnp-b raw-GWA-data.hapmap3r2.pruned.pedind -o raw-GWA-data.hapmap3r2.pruned.pca -p raw-GWA-data.hapmap3r2.pruned.plot -e raw-GWA-data.hapmap3r2.pruned.eval -l raw-GWA-data.hapmap3r2.pruned.log -k 2 -t 2 -w pca-populations.txt`'.

18| Create a scatter diagram of the first two principal components, including all individuals in the file `raw-GWA-data.hapmap3r2.pruned.pca.evec` (the first and second principal components are columns 2 and 3, respectively). Use the data in column 4 to color the points according to sample origin. An R script for creating this plot (`plot-pca-results.Rscript`) is provided (although any standard graphing software can be used).

19| Derive PC1 and PC2 thresholds so that only individuals who match the given ancestral population are included. For populations of European descent, this will be either the CEU or TSI HapMap3 individuals. Here, we chose to exclude all

individuals with a second principal component score less than 0.072. Write the FID and IID of these individuals to a file called 'fail-ancestry-QC.txt'.

▲ CRITICAL STEP Choosing which thresholds to apply (and thus which individuals to remove) is not a straightforward process. The key is to remove those individuals with greatly divergent ancestry, as these samples introduce the most bias to the study. Identification of more fine-scale ancestry can be conducted by using less divergent reference samples (e.g., within Europe, stratification could be identified using the CEU, TSI (Italian), GBR (British), FIN (Finnish) and IBS (Iberian) samples from the 1,000 Genomes Project (http://www.1000genomes.org/)). Robust identification of fine-scale population structure often requires the construction of many (2–10) principal components.

Removal of all individuals failing QC ● TIMING ~5 min

20| At the Unix prompt, type 'cat fail-* | sort -k1 | uniq > fail-qc-inds.txt' to concatenate all files listing individuals who fail the previous QC steps into a single file.

21| The file fail-qc-inds.txt should now contain a list of unique individuals failing the previous QC steps. To remove them from the data set, type 'plink --bfile raw-GWA-data --remove fail-qc-inds.txt --make-bed --out clean-inds-GWA-data' at the Unix prompt.

Identification of all markers with an excessive missing data rate ● TIMING ~20 min

22| To calculate the missing genotype rate for each marker, type 'plink --bfile clean-inds-GWA-data --missing --out clean-inds-GWA-data'. The results of this analysis can be found in clean-inds-GWA-data.lmiss.

23| Plot a histogram of the missing genotype rate to identify a threshold for extreme genotype failure rate. This can be carried out using the data in column 5 of the clean-inds-GWA-data.lmiss file and any standard statistical/graphing software package. A script for creating this histogram in R is provided (lmiss-hist.Rscript). We chose a call-rate threshold of 3% (these SNPs will be removed in Step 26).

Test markers for different genotype call rates between cases and controls ● TIMING ~ 5 min

24| At the Unix prompt, type 'plink --bfile clean-inds-GWA-data --test-missing --out clean-inds-GWA-data' to test all markers for differences in call rate between cases and controls. The output of this test can be found in clean-inds-GWA-data.missing.

25| At the Unix prompt, type 'perl run-diffmiss-qc.pl clean-inds-GWA-data' to create a file named 'fail-diffmiss-qc.txt', which contains all SNPs with a significantly different ($P<0.00001$) missing data rate between cases and controls.

Removal of all markers failing QC ● TIMING ~ 5 min

26| To remove poor SNPs from further analysis and to create a clean GWA data file, type 'plink --bfile clean-inds-GWA-data --exclude fail-diffmiss-qc.txt --maf 0.01 --geno 0.05 --hwe 0.00001 --make-bed --out clean-GWA-data' at the Unix prompt. In addition to the markers failing previous QC steps, those with an MAF<0.01 and an HWE P-value < 0.00001 (in controls) are also removed.

● TIMING

Steps 1–3, Creation of BED files: ~ 20 min
Steps 4–6, Identification of individuals with discordant sex information: ~30 min
Steps 7–10, Identification of individuals with elevated missing data rates or outlying heterozygosity rate: ~30 min
Steps 11–13, Identification of duplicated or related individuals: ~4 h
Steps 14–19, Identification of individuals of divergent ancestry: ~2 h
Steps 20 and 21, Removal of all individuals failing QC: ~5 min
Steps 22 and 23, Identification of all markers with an excessive missing data rate: ~20 min
Steps 24 and 25, Test markers for different genotype call rates between cases and controls: ~5 min
Steps 26, Removal of all markers failing QC: ~5 min
Box 3, candidate-gene study: 15 min

Inexperienced analysts will typically require more time. Given the computational nature of this protocol, timing will also vary with computational resources.

? TROUBLESHOOTING

Step 15: It is likely that the merge will not complete and PLINK will terminate with the message 'ERROR: Stopping due to mis-matching SNPs -- check +/ – strand?'. Read `raw-GWA-data.hapmap3r2.log` to see this message. Because all A→T and C→G SNPs have been removed before undertaking this analysis, all SNPs that are discordant for DNA strands between the two data sets are listed in the `raw-GWA-data.hapmap3r2.pruned.missnp` file. To align the strands across the data sets and successfully complete the merge, simply repeat Step 14, including the command '`--flip raw-GWA-data.hapmap3r2.pruned.missnp`', and then repeat Step 15.

For help with the programs and websites used in this protocol, refer to the following relevant links:
PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml
SMARTPCA: http://genepath.med.harvard.edu/~reich/Software.htm
R: http://cran.r-project.org/
HapMap: http://hapmap.ncbi.nlm.nih.gov/

ANTICIPATED RESULTS

Genome-wide association studies

Identification of individuals with elevated missing data rates or outlying heterozygosity rate. Examining a plot of genotype failure rate versus heterozygosity across all individuals in a study allows one to identify samples introducing bias into a study (**Fig. 1**). We chose to exclude all individuals with a genotype failure rate ≥0.03 (**Fig. 1**, vertical dashed line) and/or a heterozygosity rate ± 3 s.d. from the mean (**Fig. 1**, horizontal dashed lines).

Identification of individuals of divergent ancestry. Principal component models were built using the CEU, CHB + JPT and YRI HapMap3 samples and then used to predict principal component scores for cases and controls (of supposed European ancestry) using smartpca.pl (**Fig. 2**). We chose to exclude 30 individuals who clustered away from the HapMap European samples (CEU). Depending on the population from which cases and controls are selected, the GWA study samples may not cluster precisely with population reference samples. Population stratification can still be reduced by removing individuals who lie away from the main cluster of GWA study samples (even if they do not cluster over a reference population). Alternatively, one can use more appropriate population reference samples in the PCA analysis.

Identification of all markers with an excessive missing data rate. Scrutinizing the distribution of the SNP call rate is one of the best ways to assess the success of the genotyping experiment (**Fig. 3**). SNPs with an excessive amount of missing genotypes should be removed to reduce false positives.

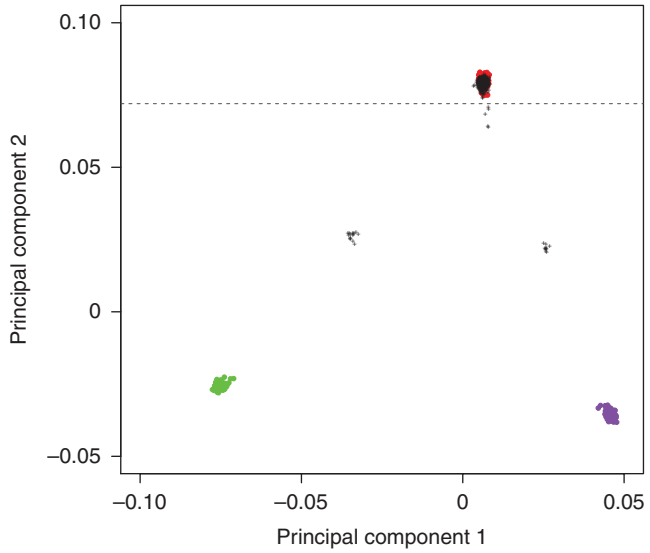

Figure 2 | Ancestry clustering based on genome-wide association data. HapMap3 reference samples are CEU (red), CHB + JPT (purple) and YRI (green). Genome-wide association samples are shown as black crosses. Cases (n = 11) and controls (n = 19) with a second principal component score less than 0.072 (gray dashed line) were marked for removal.

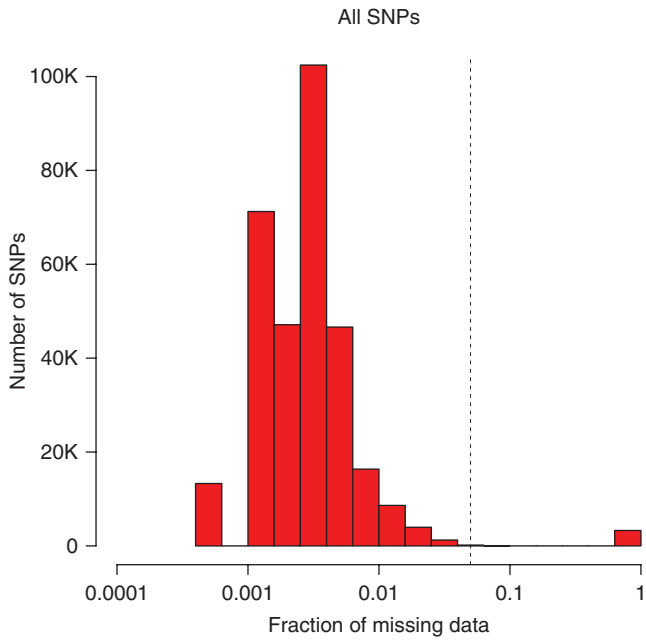

Figure 3 | Histogram of missing data rate across all individuals passing 'per-individual' quality control. The dashed vertical line represents the threshold (3%) at which SNPs were removed from further analysis because of an excess failure rate.

Candidate-gene study

Test markers for different genotype call rates between cases and controls. For the *PPARG* SNPs in the file provided, no SNPs failed this QC check.

Running additional QC steps and removal of failing markers and samples. The clean *PPARG* data files should contain 1,971 cases and 1,989 controls genotyped across 25 SNPs with a genotype call rate of 0.986.

Note: Supplementary information is available in the HTML version of this article.

ACKNOWLEDGMENTS C.A.A. was funded by the Wellcome Trust (WT91745/Z/10/Z). A.P.M. was supported by a Wellcome Trust Senior Research Fellowship. K.T.Z. was supported by a Wellcome Trust Research Career Development Fellowship.

AUTHOR CONTRIBUTIONS C.A.A. wrote the first draft of the article. C.A.A. wrote scripts and performed analyses. C.A.A., F.H.P., G.M.C., A.P.M. and K.T.Z. revised the article. C.A.A., L.R.C., A.P.M. and K.T.Z. designed the protocol.

COMPETING FINANCIAL INTERESTS The authors declare no competing financial interests.

Published online at http://www.natureprotocols.com/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Zondervan, K.T. & Cardon, L.R. Designing candidate gene and genome-wide case–control association studies. *Nat. Protoc.* **2**, 2492–2501 (2007).
2. The Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
3. Anderson, C.A. *et al.* Investigation of Crohn's disease risk loci in ulcerative colitis further defines their molecular relationship. *Gastroenterology* **136**, 396–399 (2009).
4. Teo, Y.Y. *et al.* A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**, 2741–2746 (2007).
5. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case–control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
6. Marchini, J., Howie, B., Myers, S.R., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
7. Silverberg, M.S. *et al.* Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nat. Genet.* **41**, 216–220 (2009).
8. Pompanon, F., Bonin, A., Bellemain, E. & Taberlet, P. Genotyping errors: causes, consequences and solutions. *Nat. Rev. Genet.* **6**, 847–859 (2005).
9. Price, A.L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 132–135 (2008).
10. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
11. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
12. Campbell, C.D. *et al.* Demonstrating stratification in a European American population. *Nat. Genet.* **37**, 868–872 (2005).
13. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1616–1617 (1996).
14. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
15. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
16. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
17. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
18. Fisher, S.A. *et al.* Genetic determinants of ulcerative colitis include the ECM1 locus and five loci implicated in Crohn's disease. *Nat. Genet.* **40**, 710–712 (2008).
19. Wittke-Thompson, J.K., Pluzhnikov, A. & Cox, N.J. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 967–986 (2005).
20. Meyre, D. *et al.* Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nat. Genet.* **41**, 157–159 (2009).
21. Moskvina, V., Craddock, N., Holmans, P., Owen, M.J. & O'Donovan, M.C. Effects of differential genotyping error rate on the type I error probability of case–control studies. *Hum. Hered.* **61**, 55–64 (2006).
22. Plagnol, V., Cooper, J.D., Todd, J.A. & Clayton, D.G. A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet.* **3**, e74 (2007).
23. Morris, A.P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
24. Pettersson, F.H. *et al.* Marker selection for genetic case–control association studies. *Nat. Protoc.* **4**, 743–752 (2009).
25. R Development Core Team. R: a language and environment for statistical computing. (2005).
26. Aulchenko, Y.S., Ripke, S., Isaacs, A. & van Duijn, C.M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
27. Pettersson, F., Morris, A.P., Barnes, M.R. & Cardon, L.R. Goldsurfer2 (Gs2): a comprehensive tool for the analysis and visualization of genome wide association studies. *BMC Bioinformatics* **9**, 138 (2008).
28. Pettersson, F., Jonsson, O. & Cardon, L.R. GOLDsurfer: three dimensional display of linkage disequilibrium. *Bioinformatics* **20**, 3241–3243 (2004).