H3A - Genome-Wide Association testing SOP

Phase 2 (quality control)

1. check for overall population structure
    a. between cases vs controls
    b. batch effects
    c. amongst reference populations (eg KGP)
2. relatedness

Phase 3 (disease association testing)

1. LD calculation
2. Impute missing genotypes (if applicable)
3. Association testing
4. Local ancestry

Introduction

The following document describes some recommended steps for processing genotype data from a genome-wide association study, as will be generated by many of the H3Africa projects. Although it is written with data from SNP genotyping arrays in mind, many of the steps also apply to full genome sequence and exome sequencing data. Many of the tests can be performed with the PLINK software suite, but where other software is required this will be indicated.

File format

This SOP assumes that incoming data is in the binary format used by the PLINK software suite. The PLINK binary format (hereafter referred to as bped) encodes a dataset as a set of three files, with the following suffixes to their names:

.bed: this is the binary genotype data, stored as 2 bits per genotype per sample
.bim: a text file containing marker information, one line per marker, in genomic order
.fam: a text file containing sample pedigree information

Strand errors

As a byproduct of the genotyping technology, and the annotation data that accompanies the chips, SNP data from both Illumina and Affymetrix platforms may be reported as the allele on either the “forward” or “reverse” strand. Although the information to orient these calls properly is contained within the annotation data, conversion from the native format to PLINK format can be tricky, and errors will not be evident in “ambiguous” A-T/G-C SNPs. Therefore it is prudent to check that the alleles reported in the bim file match the known alleles in dbSNP. Another easy check is to identify the control samples that are often included in the dataset (eg samples from HapMap) and compare these to their known genotypes.

Sample quality control

Before association testing can be done, some sample QC steps should be taken to minimise confounding factors. Some of these filters would have been applied at the genotype calling phase, but it is worth confirming them if the genotype calling was not done internally.

Sample call rate
Samples with poor call-rates should have been removed at the genotype calling stage, since this is a strong indicator of poor sample quality (perhaps due to problems in collection, processing or hybridisation) and will adversely affect the intensity clusters used to call genotypes, affecting the calling of all samples. A typical filter is to remove samples with a call rate lower than 98%. This filter should be run after the marker call-rate filter mentioned elsewhere in this document, to reduce the effect of poorly performing probes on a sample's call rate
Samples with a low call rate can be removed with PLINK's --mind option.

Identity errors
The sex of an individual can be determined from the genotype data, and should be compared to the declared sex recorded in the phenotype records. Samples with a mismatched sex call should be investigated to determine whether this points to errors in other samples.
Checking sex of a sample can be done with PLINK's --check-sex option.

Samples with chromosomal abnormalities
Chromosomal abnormalities such as aneuploidy or long stretches of homozygosity should be tested for and the causes identified. These cases should be examined with a geneticist to determined possible causes and decide whether the sample should be removed.

Control samples
To assist in the estimation of genotyping accuracy, samples with well known genotypes (for example from the HapMap project) are often included in the genotyping pipeline. These genotypes should be compared to the known data, and can be used as a measure of accuracy.

Another quality control strategy is to include replicates from the group under study. These should be compared with each other and markers that are not concordant should be filtered.

Concordance between two samples can be measured with PLINK's --bmerge and --merge-mode options

Reported relationships between samples
Another check of sample identity is to confirm that relatedness between individuals matches their reported relationships. This can serve to highlight:

- labelling issues
- average degree of relatedness between individuals across the whole population
- non-paternity

- duplicate samples not labelled as such

Pairwise relatedness can be computed using PLINK’s --genome option

Marker quality control

Marker call rate

On arrays with hundreds of thousands of probes, some markers can be expected to perform poorly. These should be removed from the dataset because it is likely that even the samples that were “successfully” called are inaccurate due to poor clustering of the hybridisation intensities. A typical filter is to remove markers with a call rate lower than 98%
Samples with a low call rate can be removed with PLINK’s --geno option.

Minor allele frequency

Removal of SNPs with a low MAF (minor allele frequency) is often recommended because they have low statistical power, increase the multiple test correction required, and might correlate with poor calling. A typical threshold is 0.01, but this is highly dependent on the size of the study.

Power calculation

Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium should be calculated for each SNP, and flagged since this can point to several factors such as population stratification, poor genotype calling or even true associations with the study traits. An in-depth examination of HWE and GWAS can be found in *Calculation and use of the Hardy-Weinberg model in association studies.*[1]
HWE can be calculated in PLINK with the --hardy option.

Batch effects

Batch effects refers to systematic differences in results that correlate with factors not under study, such as:

- sampling site
- 96-well plates
- date of sample processing

For multi-center studies (which many H3Africa studies are), phenotype standardisation should be confirmed. Systematic differences in phenotype collection can introduce batch effects which should be tested for.

These should be tested for by measuring across each batch

[1] http://www.ncbi.nlm.nih.gov/pubmed/18428419

Population stratification

Population stratification refers to structure within the sample group that is the result of systematic genetic differences between individuals that correlate with the phenotypic data. This could result in allele frequency differences that are due to ancestral proportion differences between cases and controls being mistaken for an association with the phenotype.

Two methods can be used to examine population substructure:

STRUCTURE/Admixture

Allele frequency based methods, such as STRUCTURE.

Add known reference populations such as hapmap pops
Determine most likely K
Look for:
 - differences between possible batch
 - samples that cluster with the wrong group
 - look for individuals that cluster with the wrong population, or have admixture pattern different from others other samples in the same group

Eigensoft

PCA based methods, such as Eigensoft. Eigensoft’s smartpca, and examine the top 10 eigenvectors. If any of these eigenvectors display segmentation based on a potential confounder, these factors should be examine further, and possibly corrected for. Individuals that differ significantly from the rest in terms of cluster membership should be investigated for possible removal. A common threshold in Eigensoft is to remove individuals that are outliers by more than 6 standard deviations.

Both these methods should be used on the data in an exploratory fashion, to try identify possible confounding factors such as:

- case/control status. One of the assumptions in GWAS is that overall, the case and control groups have the same genetic background. If stratification is different between the two groups, this could lead to false positives
- batch effects. Check for homogeneity across 96-well plates, collection centers etc. A common test is to perform PCA on the dataset, and label the samples according to these phenotypes
- other collected data. e.g. socioeconomic status
- structure in comparison to reference populations such as those from HapMap/1000 genomes

Outliers identified in PCA analysis should also be removed from further analysis, since these could indicate poor samples, mislabelled individuals

Association testing

Single locus tests

The most basic test of association is the single locus test.

1. LD calculation
2. Impute missing genotypes
3. Association testing
   a. Use plink to do a simple association testing
   b. use EMMAX which adjusts for population stratification/relatedness

Adjustment for covariates

Replication

The relatively low power of most GWAS designs means that a substantial number of false positives can be expected to be generated, so validation via replication in an independent population is an important part of most studies.

A common strategy is to genotype a subset of the samples on a high coverage array, and then follow up with targeted genotyping of markers that appear interesting, on a larger number of samples

Meta analyses

References: