ELSA GWAS QC & Imputation Analysis

Report prepared by Chrissy h Roberts : chrissyhroberts@yahoo.co.uk

30/09/2017

Contents

Overview

This document describes steps that have been taken to explore, catalogue and impute data in the version of the ELSA genome wide genotyping data set that was provided via the European Genome/Phenome Archive (EGA) website. The versions of the data used here were originally downloaded on December 13th 2013, but were still current on April 20th 2017. All the analyses and explorations are based on these files. For reference, we provide MD5 checksums for these files. By checking that the versions of the files you download from EGA have the same MD5 checksums, you can ensure that any analysis you perform will be based on the exact same files.

If you plan to use the Imputed data files described in this document then you should do so only

There are some guides to the various files that you will need to use in your analysis. These can be found at the end of this document. Please pay close attention to what they have to say or you may find that you could end up making claims about genetic associations that are not robust.

Files used and MD5 checksums

ega-box-163_ForwardStrand_excREL.fam (MD5 Checksum 812997da421fec29f9760a6273ebe570)
ega-box-163_ForwardStrand_excREL.bed (MD5 Checksum 503b0433bd683ce416718cd24b935606)
ega-box-163_ForwardStrand_excREL.bim (MD5 Checksum 9d58028ba89b606e3804ef47947dad6d)

Files generated during this analysis

The most practically useful things that we provide are a set of files that include imputed genotypes. These data are derived from a data set that we first subjected to rigorous and rather conservative quality control screening and filtering. We had to remove a few SNPs (files accompanying this document list which ones) to ensure that the imputation process was robust. We have not however removed any individuals from the data, although there are a number of individuals who for one reason or another do need to be removed.

During the analyses described in this document we have generated and distributed a number of new files which may be useful to end-users, especially when performing initial cleaning and filtering steps.
We make some recommendations for Single Nucleotide Polymorphisms (SNPs) [also referred to within this document as 'features'] and for specimens (from individual ELSA participants) that we believe should be removed before genetic association tests are performed.

For ease of use we provide a set of files that can be used with the industry standard genome-wide association screening (GWAS) analysis platform PLINK (https://www.cog-genomics.org/plink/1.9) to extract a tidy and well quality controlled data set for downstream analysis.

Throughout these analyses we have used PLINK v1.90 beta and R v3.3.2. This document includes extracts from the PLINK log files that were generated during the analyses. All the steps should be replicable by running PLINK again with the parameters set out in those extracts.

Identify duplicated features

The PLINK files contain a number of Illumina QC features (labelled Chr0 in the .bim file) and some features that appear twice due to naming conflicts. We identified all features with duplicated coordinates and allele coding, then removed one instance of each, preferring any identifiers where an Illumina 'KGP' coding was present in addition to a standard 'rs' identifier. The Illumina KGP identifiers can be updated with rs identifiers using PLINK's various –update commands and an appropriate list of SNP rsids and coordinates from UCSC genome browser (https://genome.ucsc.edu/)

Requirements

- ega-box-163_ForwardStrand_excREL.bim
- ega-box-163_ForwardStrand_excREL.fam
- ega-box-163_ForwardStrand_excREL.bed

The initial data set contains a number of features that have duplicated coordinates. Whilst this is in some cases the result of two SNPs at the same chromosomal position, it is a nuisance factor that most analyses can't control for. In any case the number of duplicated coordinates is extremely low and in many cases the upstream data prep strategies may have led to misnaming or erroneous assignment of rsids to such positions. The conservative strategy is simply to remove one SNP at each duplicated locus.

Because the Illumina Omni chip had a set of prototype SNPs for which there were no rsids yet assigned, the naming conventions on the duplicated features in this data set

The list of duplicated features has been written to a file

> Exclusions_SNPs_Duplicated_features.txt

The number of unique (FALSE) and duplicate (TRUE) feature locations is as follows

```
##   FALSE    TRUE
## 2298348   10670
```

Instructions to Remove duplicates

Use the following command to replicate this step in your own analysis.

> PLINK –noweb –bfile ega-box-163_ForwardStrand_excREL –exclude Exclusions_SNPs_Duplicated_features.txt –make-bed –out ega_data_no_dup

Identify and remove strand flips using SNPFlip software

The file names that are downloaded from EGA indicate that the PLINK files are aligned to a 'forward' strand, but it is clear from the outset that a number of SNPs were aligned to the other strand. SNP-FLIP (https://github.com/mdshw5/snp-flip) was used to compare the reference assembly of the 1000 Genomes Project, GRCh37 FASTA (human_g1k_v37.fasta.gz) [http://www.internationalgenome.org/category/grch37/] to the ELSA data files. During this analysis it was neccessary to recode the X chromosome designations in the original EGA files from the PLINK style '23' to 'X' in order to match the fasta files from the 1000G project.

Please note that QC, filtering, imputation and analysis of non-autosomes is a complex and difficult process that is outside the scope of the analyses described in this document. Sex chromosome coding has been changed only to ensure compatibility with the FASTA reference. We have neither considered the Y chromosome, nor the pseudoautosomal regions (PAR) of Y, nor Mitochondrial genome. We have also removed the illumina QC features (designated Chr 0) from the bim file.

Requirements

- SNP-FLIP (https://github.com/mdshw5/snp-flip)
- GRCh37 FASTA file (871MB) from 1000G (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/human_g1k_v37.fasta.gz)
- HumanOmni2.5-4v1_D-b37.strand (http://www.well.ox.ac.uk/~wrayner/strand/)
- 001_ega-box-163_ForwardStrand_excREL_recoded_sex_chromosomes
- ega_data_no_dup.bim
- ega_data_no_dup.log
- ega_data_no_dup.bed
- ega_data_no_dup.fam
- human_g1k_v37.fasta

Instructions to replicate this analysis by performing SNPFLIP analysis

snpflip -b ega_data_no_dup_recoded_sex_chromosomes.bim -f human_g1k_v37.fasta -o snpflip_output_initial

SNPFLIP results

SNPFLIP compares the 'orientation' of the features in the data set to those in the reference files. Strand flips are in general fairly harmless, although some people don't like them to appear in the data and some software aren't smart enough to consider the orientation before performing tests and analysis.
SNPFLIP reports data on each feature, classifying them in to three bins : 'ambiguous', 'forward' and 'reverse'.

The ambiguous types (AT, TA) look the same in either orientation, so they have little value in our data set and should be removed. The forward orientation is preferred by most users.

The results of the SNPFLIP analysis were as follows

```
##
Read 34.3% of 2303099 rows
Read 55.6% of 2303099 rows
Read 72.5% of 2303099 rows
Read 92.9% of 2303099 rows
Read 2303099 rows and 9 (of 9) columns from 0.100 GB file in 00:00:07

## ambiguous   forward   reverse
##     72916   2111599    118584
```

Instructions to exclude ambiguous features and flip those on reverse strand

> PLINK –noweb –bfile ega_data_no_dup –flip snpflip_output_initial.reverse2 –exclude snpflip_output_initial.ambiguous2 –make-bed –out ega_data_flipped

Change reference allele order to match the 1000G file

When files are generated in PLINK, the software automatically sets the major allele as the reference allele. Some people are not happy with this and would prefer their data to use the same reference alleles as the 1000 Genomes reference files, against which imputation will be performed. This really doesn't make any practical difference, but we have realigned all the data for the sake of pleasing purists.

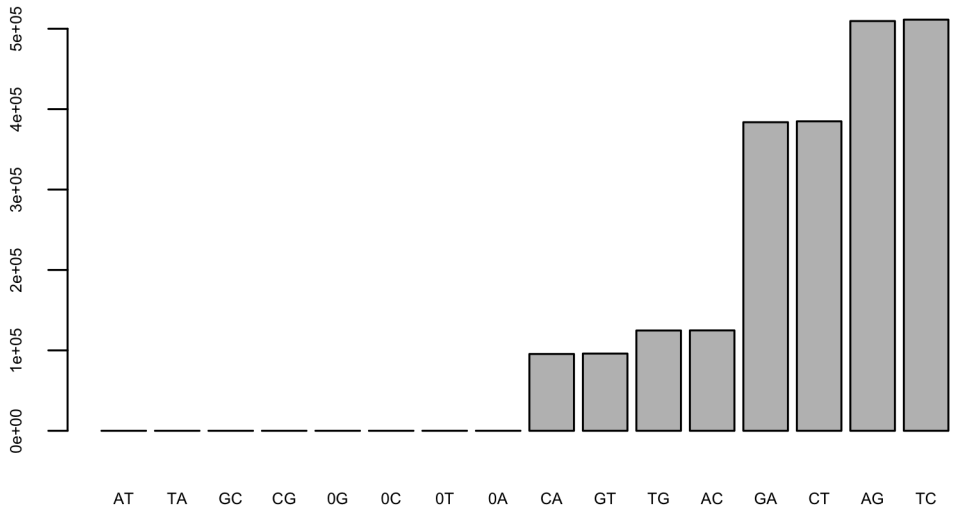


Figure : Number of features in each genotype configuration

Check association results are common when test is with (a) original and (b) flipped data

For this final test of the quality of the new data set and to demonstrate that the manipulations we had performed had not affected any structural changes in the data, we generated a simulated set of null phenotype data and performed an association test using both the original "ega-box-163_ForwardStrand_excREL" and new data sets. The results of this comparison should and did show that identical test statisticswere obtained for all SNPs in both analyses.

Flipping the strands and manipulating the reference alleles in the manner described above has had no effect on the association testing. The process did not alter genotypes in any deleterious manner.

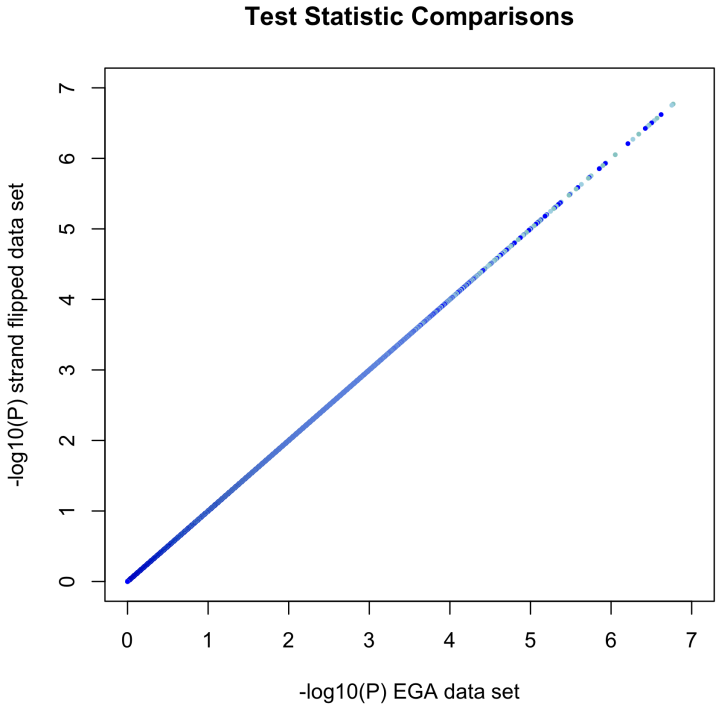


Figure : Comparison of test statistic (P value) from simulated data using (a) initial and (b) flipped data

Impute genders and perform sex checks

At this stage we initiate a series of data quality tests, starting with a test of whether the self-described gender identity from the main ELSA database (variable "dhsex") matches the empirically determined (genetic) gender, as indicated by the F coefficient on the X chromosome. This test essentially measures the extent of homozygosity on the X chromosome. The majority of males have a single X chromosome and the average male should therefore approach 100% homozygosity at all loci on X. Most females have two X chromosomes and the average female should be heterozygous for many features across the X chromosome. In practice, males usually have an F coefficient on the X chromosome of more than 0.8, whilst females usually have an F coefficient on the X chromosome of less than 0.2. Some individuals may have genetically more complex chromosome arrangements and may fall outside the 'normal' ranges. Low quality genotyping can also be a cause of unusual values for this statistic.

Instructions to replicate this analysis

> plink –noweb –bfile ega_data_flipped_ref_reset –keep-allele-order –impute-sex –make-bed –out 001_elsa

Results of gender analysis

```
## Number of 0 = unknown 1 = male 2 = female genotypes
```

```
##    0    1    2
##   11 3426 3975
```

```
## Number of gender mismatches
```

```
## MATCH  NA's
##  5758  1654
```

```
## Comparison of imputed gender (x) against recorded gender (y) | 0 = unknown 1 = male 2 = female
```

```
##
##        0    1    2
##   1    0 2635   13
##   2    9   18 3123
```

Comparison of empirical gender test against self-reported gender identity

The file "Exclusions_Specimens_Empirical_Sex_Test.txt" can be used to exclude those with unusual F coefficient values.

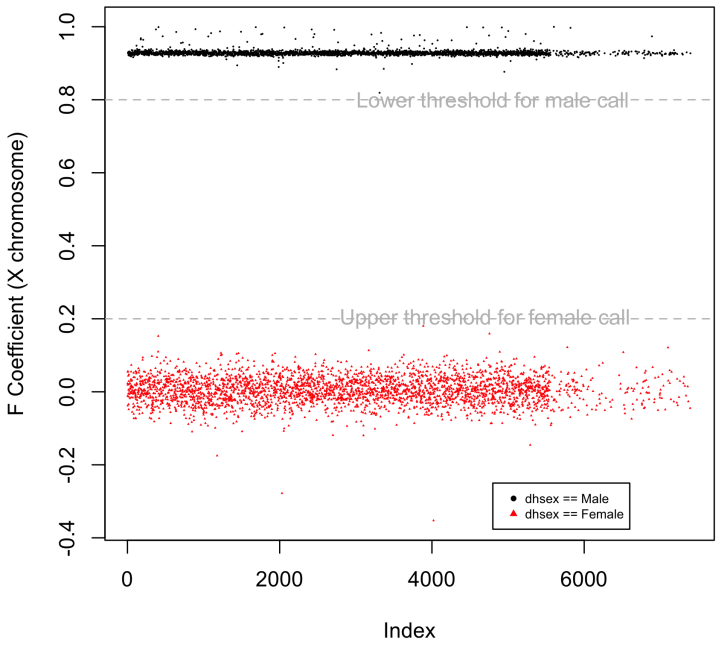


Figure : Comparison of empirical gender estimation to self-described gender identity

Outlier analysis and filtering.

We now apply a number of approaches to identifying both SNPs and features that we would consider outliers and which should potentially be removed from the analysis

Filter specimens on 2.5% missing data and 1.96 SD of mean heterozygosity (inbreeding coefficient)

In this analysis, we identify specimens in which there was a surfeit of missing data, which could indicate that those DNA specimens did not perform well on the genotyping array and/or on the software basecaller. We also identify specimens where the inbreeding coefficient (F statistic on autosomes) exceeded the 95% confidence interval for the population.

List of all F scores and missingness data written to file
>Exclusions_Specimens_missingness_and_f_score.txt
List of specimens to be excluded written to file
>Exclusions_Specimens_missingness_and_f_score.txt

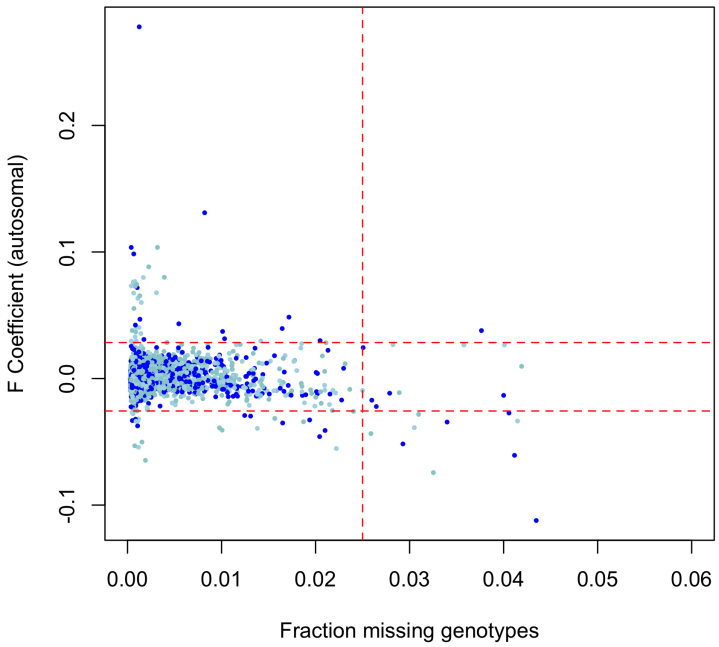


Figure : Missing data (% of data for individual specimens) and mean heterozygosity (F score). Thresholds at 2.5% missingness & F score ± 1.96 SD of mean F

```
## Results of heterozygosity/missingness test
```

```
##    Mode   FALSE    TRUE    NA's
## logical    7323      89       0
```

Filter features (SNPs) on missingness

The previous analysis identified individual specimens in which the data included large amounts of missingness and/or excessive heterozygosity. This analysis is an equivalent test for SNPs. In this analysis we identify how many SNPs would be retained in the data set if the panel was trimmed to include only SNPs with missing proportions of various sizes.

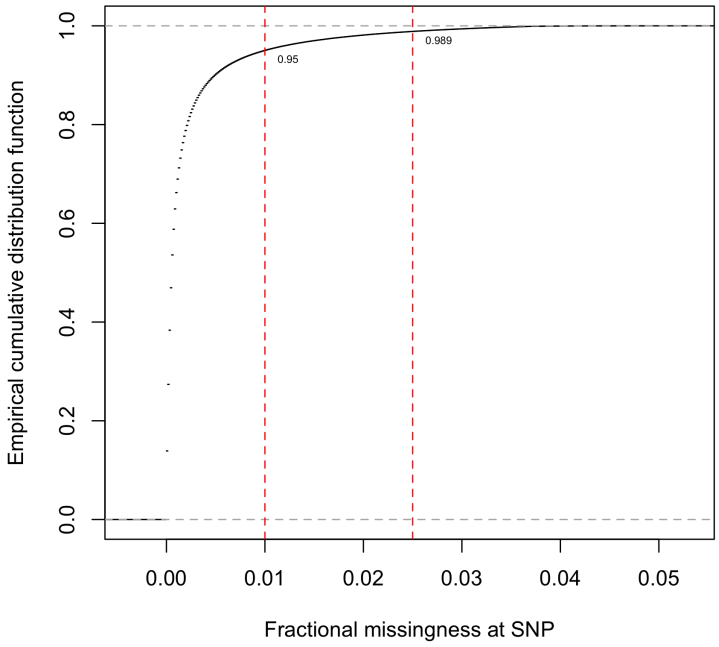


Figure : Fraction of SNPs retained when filtering at n % missingness

Check allele frequencies

In this analysis we simply check the distribution of minor allele frequencies among the SNPs in the data set. Around 35% of the data had MAF values below 1% and for basic GWAS association testing we would recommend trimming these SNPs out of the data. Imputation, rare event, burden tests and pathways analysis all benefit from the presence of low frequency alleles, so if planning to perform analyses such as this, do not remove these SNPs.

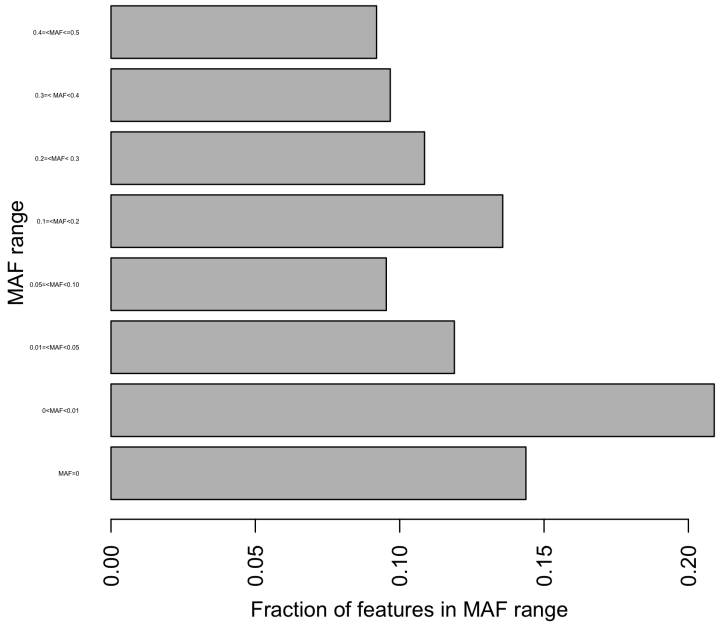


Figure : Distribution of minor allele frequencies

Identity by State (IBS) Analysis

These tests compare the amount of sharing in genome-wide genotypes between and within groups and pairs of specimens. For instance two individuals with the locus 1 genotype AA & TT share zero alleles and have an IBS score of 0 for locus 1. Other pairs might have AA & AT (IBS = 1) or TT & AT (IBS = 1), AA & AA (IBS = 2) and so on. These analyses look at average genome wide proportions of IBS scores of 0, 1 & 2

In the first such test, every individual was compared to every other individual and a score (PI^ or PI-HAT) was generated. Increasing values of PI^ indicate closer relationships between the given pair of individuals. An individual with a high average PI^ score appears to be closely related to many other individuals in the data set. Whilst some individuals may really be at the hubs of a cryptic pedigree structure in the data, this is less likely than some kind of erroneous genotyping or contamination issue with the specimen. Anyone identified as an outlier through high average PI^ score should be removed from all analyses.

To identify outliers in a robust way, we firstly generated the average PI^ scores, then performed a gaussian finite mixture model fitted by the EM algorithm to find the outliers.

A small number (n = 24) of individuals were identified as outliers during this analysis through membership of cluster 5 (See figure)

Per sample average IBS PI^ scores have been written to a file >IBS_average_PIHAT.txt

```
## best model describing clustering of specimens for outlier detection

## -------------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## -------------------------------------------------------
##
## Mclust V (univariate, unequal variance) model with 5 components:
##
##  log.likelihood    n df      BIC     ICL
##         35167.9 7412 14 70211.05 64430.1
##
## Clustering table:
##    1    2    3    4    5
## 1638 1886 2565 1299   24

## pdf
##   2
```

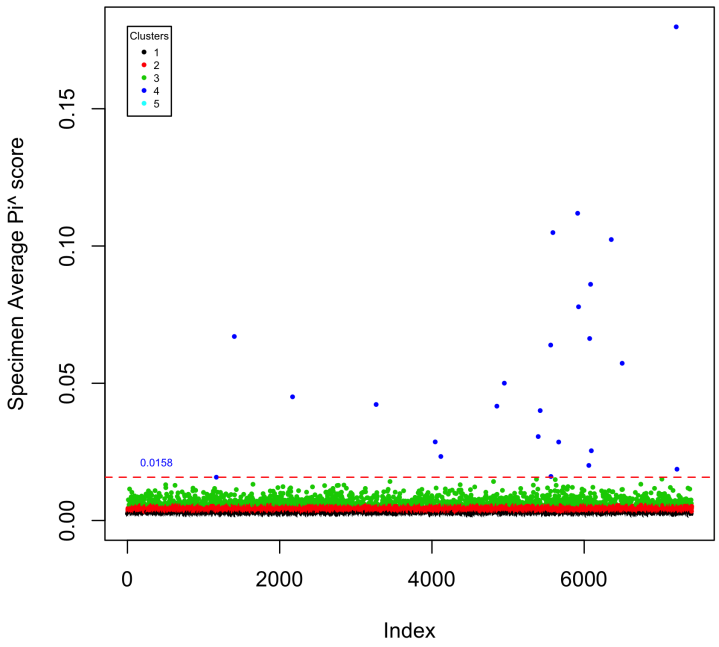


Figure : Average PIˆ value for each individual. Values estimated from pairwise IBS analysis

Nearest neigbour analysis

It can be useful to use a nearest neighbour approach to finding outliers. For each individual (the proband) we rank all other individuals according to genome-wide IBS, then ask whether their nearest neighbour is more close or more distant in IBS terms than is the average nearest neighbour in the total data set. By comparing the proband's nearest neighbour to the Z distribution of the total data, we can isolate individuals with Z scores outlying the rest of the population. By extending this analysis to small groups of specimens (i.e. asking whether the proband and their 2,3 or 4 nearest neighbours are all outliers from the population) we can find small sub-populations that should be removed. Specimens where Z > -4.0 for any of the first five nearest neighbours were considered to be outliers and were flagged for removal.

A list of specimens failng nearest neighbour test written to file : >Exclusions_Specimens_nearest_neighbour_test.txt

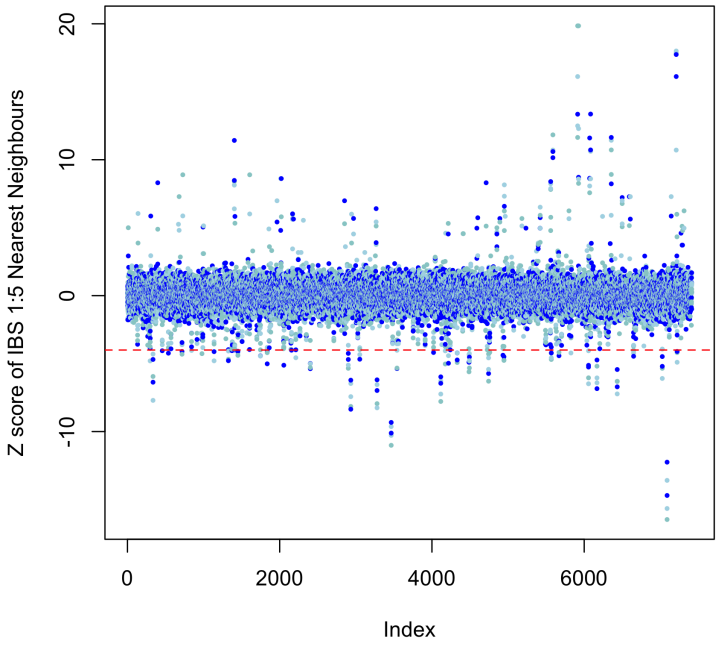


Figure : Z scores of the first to fifth nearest neighbour analysis

Hardy Weinberg Equilibrium tests

The Hardy Weinberg principle is a very important concept in genetics, but is most commonly used in association testing as a tool for discovering SNPs where the genotyping or base-calling have failed. This is because the most simple explanation for SNPs not being in Hardy Weinberg Equilibrium is erroneous genotyping data. Ordinarily we would perform these tests on the controls in a case/control setting, but for this analysis we simply tested the entire sample and assessed how many SNPs would be retained if we filtered the data set at different threshold values of the HWE test statistic.

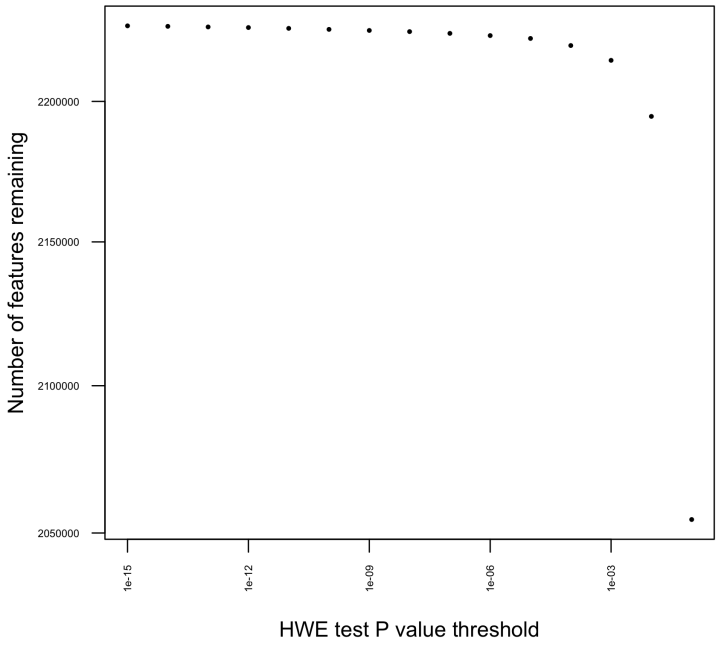


Figure : Number of features remaining if filtering at different thresholds for Hardy Weinberg Equilibrium

Population Structure Analysis

IBS kinship tests

The pairwise IBS analysis makes it possible to chart the data in such a way that very closely related (in terms of familial relationships) pairs can be identified. By charting the relationship between the proportion of IBS=0 (Z0) and the proportion of IBS=1 (Z1) you can quickly check for potentially problematic kinship structures in the data. In lay terms this means that you can find siblings, duplicate specimens, twins, parent-offspring pairs and so on.

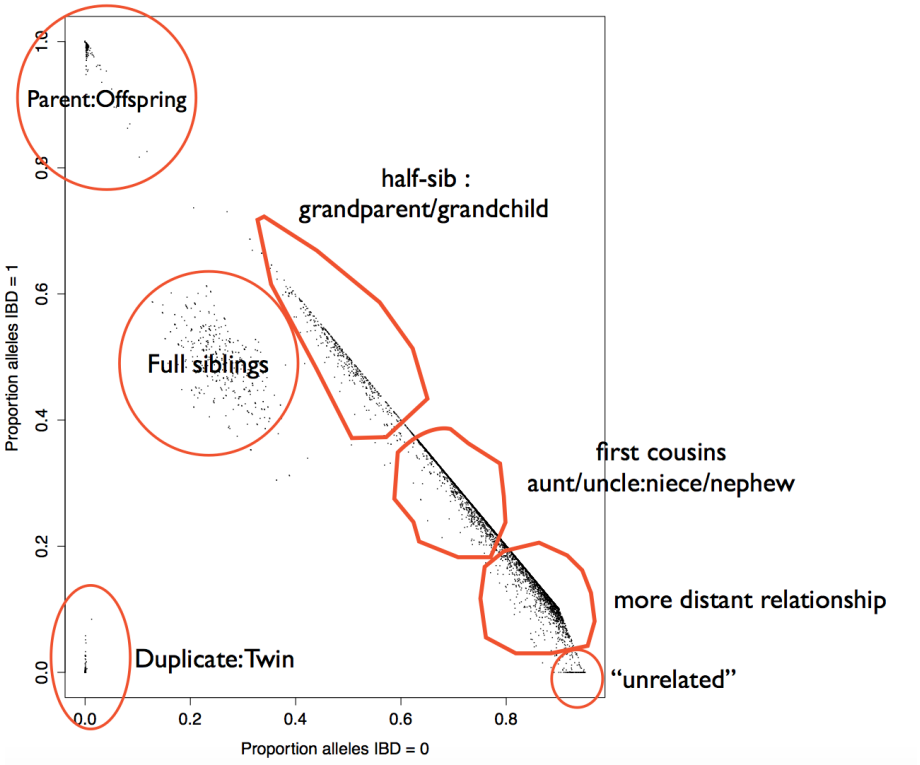


Figure : Example data showing a specimen set with extensive population structures. Each point is a pair of individuals. The position on the chart indicates Z0 & Z1 values for the pair. Ellipses indicate types of relationships that would be indicated by those values of Z0 & Z1

The ELSA data appears to have very limited kinship, with a small number of half-siblings (presumed from age of cohort not to be grandparent/grandchild relationships), cousins, aunts/uncles etc. present in the sample. This is unlikely to be substabtial enough structure to cause problems during the analysis, but mixed model association tests such as GEMMA (https://github.com/genetics-statistics/GEMMA) and EMMAX (http://genetics.cs.ucla.edu/emmax/) are able to compensate for such structure and also correct for kinship.

ELSA kinship matrix

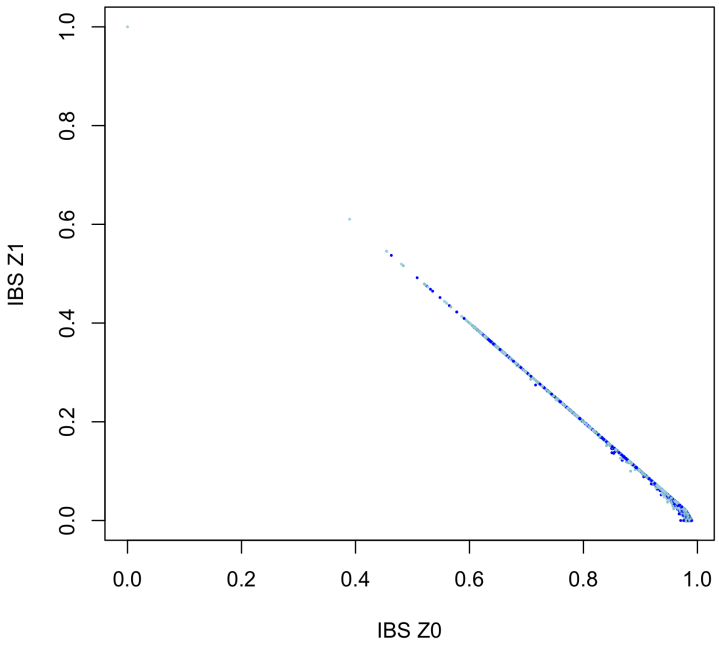


Figure : Kinship among participants. IBS sharing at (x) zero and (y) one allele. There is no evidence for siblings or parent-offspring pairs. Some second degree, third degree and more distant familial relationships were identified

Principal Components Analysis (ELSA data)

One very important test for outliers is to assess whether there is evidence for any population structure in the data at the whole genome level. Principal Components Analysis (PCA) is a convenient way to look for clustering in the data using a visualisation approach. Population structured data would present as several clusters of specimens, for instance representing people of different ethnic origins.

Figure : Principal Components Analysis (Zoom level 1)

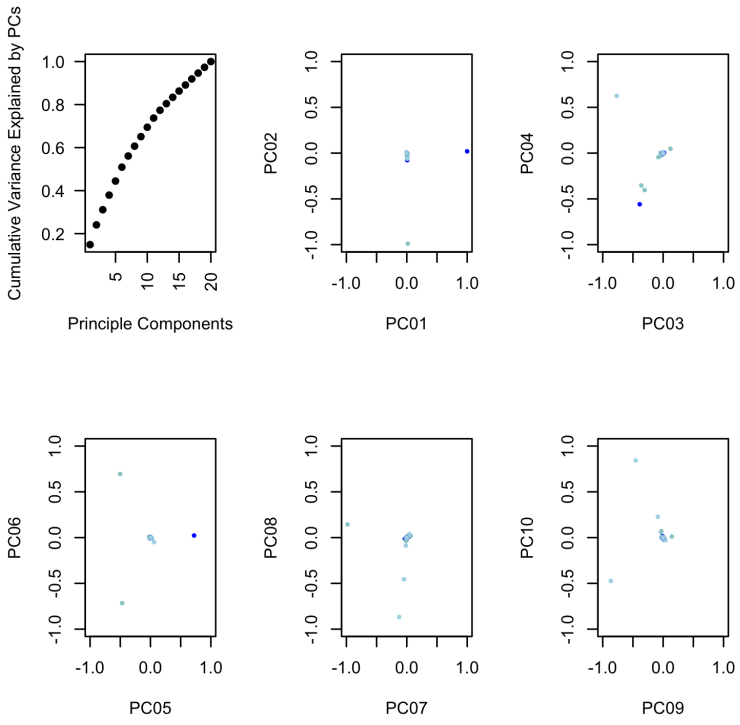


Figure : Principal Components Analysis (Zoom level 2)

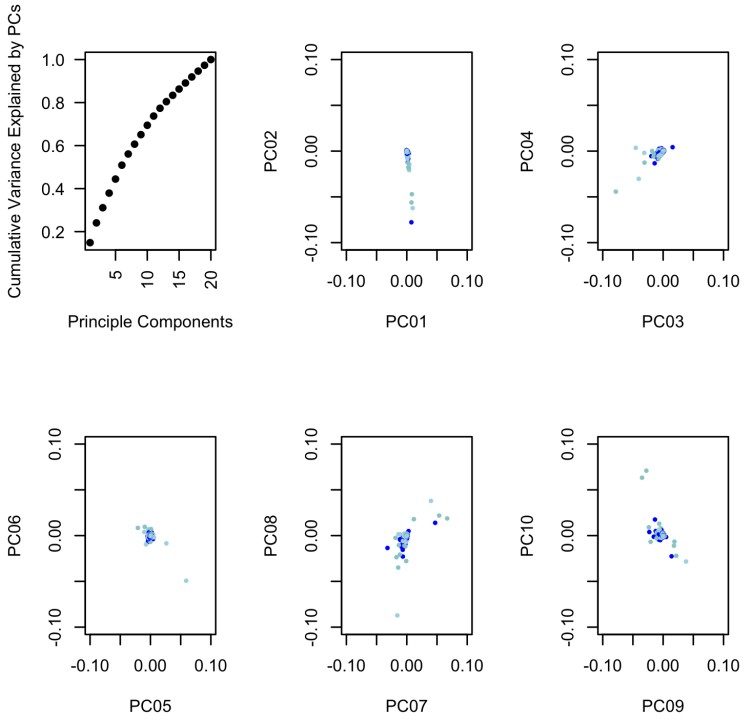
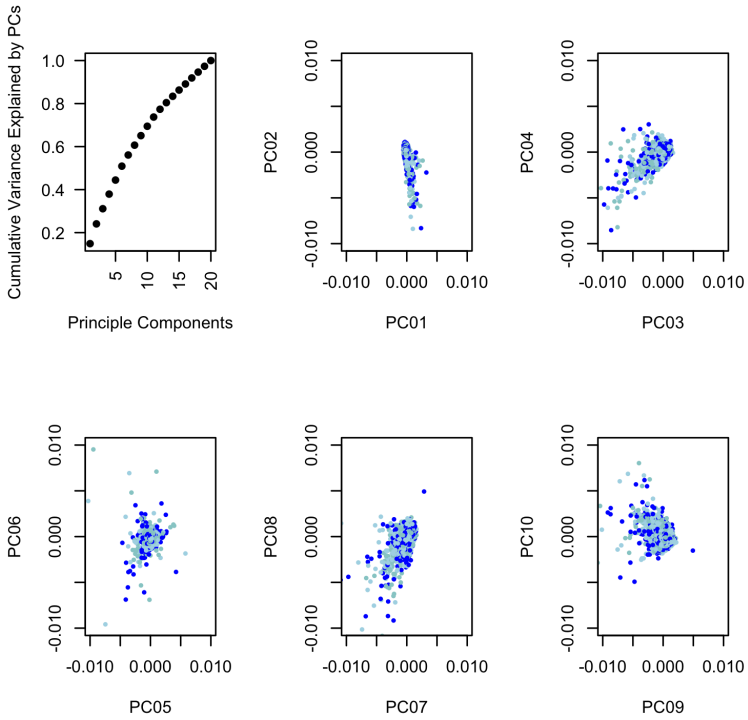


Figure : Principal Components Analysis (Zoom level 3)



Whilst there are a few individuals (see zoom level 1 figure) who appear to outlie the general population (many of whom will have been identified as outliers in the nearest neighbour analysis), there is very little evidence for any major population structure in this data set (see zoom level 3 figure). Anecdotes from the ELSA management team suggest that the ELSA cohort is an almost homogenous set of North West European Caucasians and there is nothing in this data to suggest that the sample has a mixed ethnic origin.

Principal Components Analysis (ELSA data merged with 1000G superpopulation data)

To confirm not only that the ELSA sample is an homogenous sample from a single ethnic population, but also to demonstrate that they reflect the European Caucasoid population, we compared the genomewide genotypes to those available as part of the 1000 genomes study. The two data sets were merged, retaining SNPs present in both data sets. These were then trimmed for a representative set of SNPs with low linkage disequilibrium. PCA was then performed using the 1000G superpopulation data.

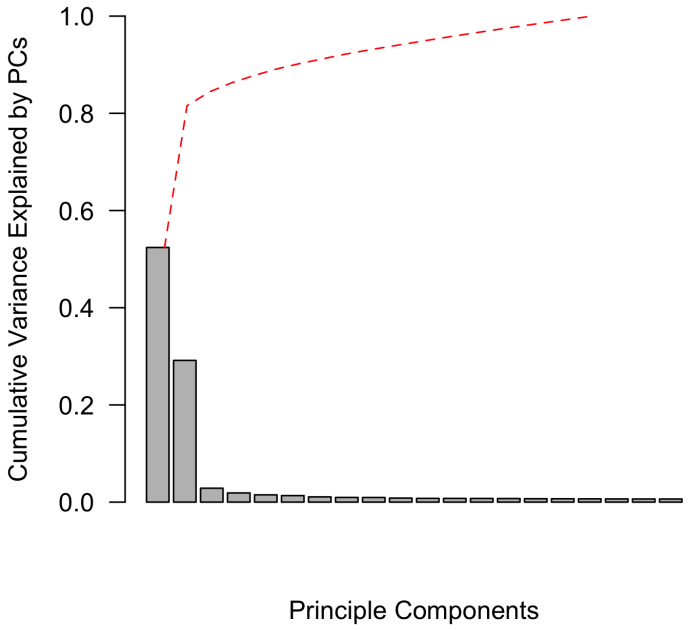


Figure : Variance Explained by PCs in combined 1000G/ELSA data. Cumulative Variance shown by red dashed line

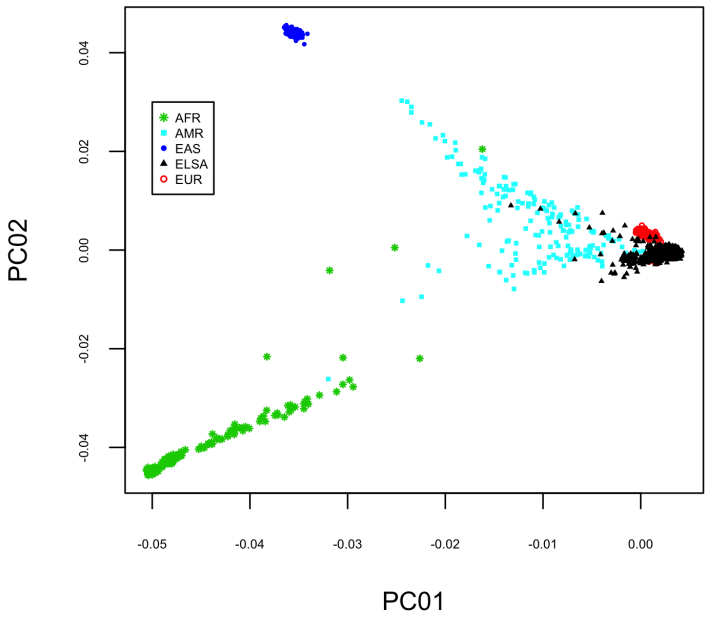


Figure : 1000 Genomes Superpopulations and ELSA (PCs 1 & 2)

This chart shows that the ELSA cohort clusters with the 1000G European population. Taken together our analysis of kinship and genomewide variance lead us to confidently conclude that genetic association studies can be carried out in the ELSA data set without the need to perform corrections for structure in the data. In lay terms this means that for most purposes it will be acceptable to use the standard PLINK association tests to perform the testing.

Imputation (Indirect genotyping)

We performed filtering on SNPs but not specimens before going on to use two well described tools, Shapeit (http://www.shapeit.fr) and IMPUTE2 (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) to impute the data sets.

Prephasing was performed with Shapeit (Delaneau et al Nat. Methods 9, 179–81 (2012) & O'Connell et al (PLoS Genet. 10, e1004234 (2014)) referencing haplotype data from HapMap Phase II, build 37. The MCMC algorithm used 35 iterations [7 Burnin iterations plus one run of 8 Pruning iterations and 20 Main iterations].

Imputation was performed with IMPUTE2 used reference data from 1092 samples included in the worldwide 1000 Genomes phase I data set. Imputation was carried out as described by Howie et al. (Nat. Genet. 44, 955–9 (2012)) in 1 MB intervals with a 20 kb overlap.

The imputed data set can be obtained from the ELSA project team https://www.elsa-project.ac.uk.

Number of SNPs in the initial data set

```
## Number of features on each chromosome (n)

## Warning: NAs introduced by coercion
```

Chromosome	# SNPs
1	178865
2	189099
3	159836
4	149421
5	141997
6	150448
7	125607
8	122558
9	100150
10	116166
11	113044
12	109852
13	81576
14	74675
15	70327
16	74068
17	64026
18	66885
19	45782
20	54811
21	31239
22	31854
23	50813

Number of SNPs in the imputed data set.

Chr	# Imp. SNPs	# MAF>1e-7	# info > 0.8	# type 2 SNPs	% concord > 0.95
1	6412190	2083048	1303843	159164	0.9726823
2	7129110	2243158	1432198	172630	0.9765684
3	5872561	1886001	1225760	144895	0.9757549
4	5772783	1900475	1231053	135543	0.9763839
5	5304859	1687618	1094342	126681	0.9758685
6	5061271	1698798	1136977	134924	0.9771797
7	4749399	1547998	986942	114460	0.9745850
8	4629656	1481540	951288	111577	0.9762406
9	3582001	1168717	738955	91449	0.9704097
10	4020491	1319079	857574	105740	0.9748912
11	4074014	1311719	846840	102753	0.9746187
12	3896094	1260300	806470	99280	0.9733380
13	2877819	966413	626480	74170	0.9733450
14	2673721	878280	551707	67866	0.9721068
15	2435545	781054	480924	63871	0.9664167
16	2718027	866196	520798	66820	0.9655941
17	2345721	773974	455938	57965	0.9586992
18	2282897	756870	476236	61189	0.9659579
19	1846241	632766	370727	40914	0.9512636
20	1825170	607058	371645	50253	0.9634649
21	1106479	375717	227751	28349	0.9617976
22	1111880	384153	221126	29051	0.9546315

Total number of imputed SNPs with MAF > 0.0000001 and Info > 0.8

```
## 16915574
```

Total Number of usable SNPs : i.e. those with info >= 0.8

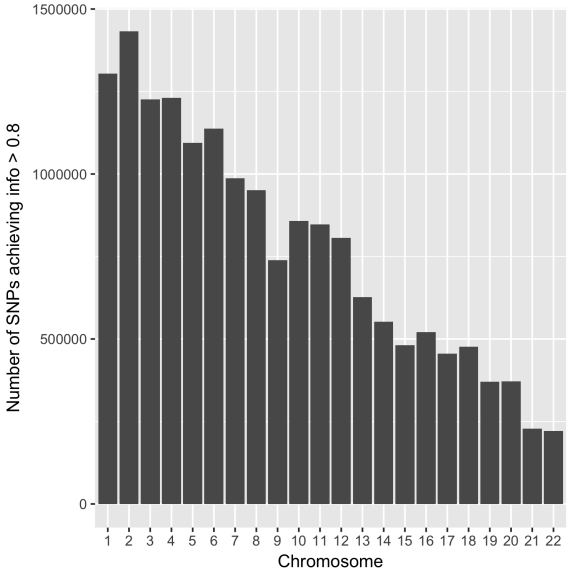


Figure : Total Number of SNPs on each chromosome with info > 0.8

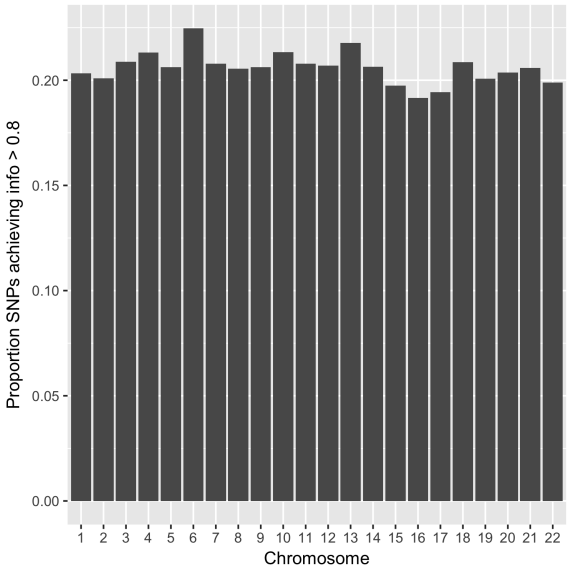


Figure : Proportion of imputed SNPs on each chromosome with info > 0.8

Proportion of type two SNPs reaching 95% concordance score in comparison of imputed to directly genotyped SNPs

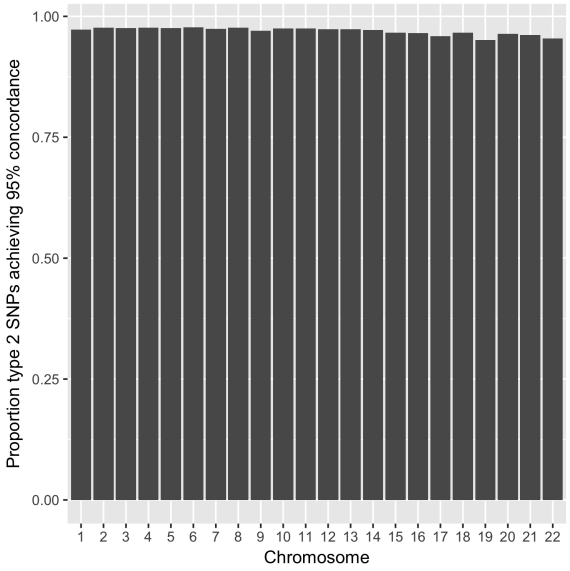


Figure : Proportion of type 2 SNPs on each chromosome achieving 95% concordance

Files supporting this document (Pre-Imputation)

Data set used for imputation

These are the clean files which were used to perform the imputation analysis

001_elsa.bed
001_elsa.bim
001_elsa.fam 001_elsa.log

The log file describes the data set 2230767 variants loaded from .bim file. 7412 people (0 males, 0 females, 7412 ambiguous) loaded from .fam. Total genotyping rate is 0.997955.

Exclusion lists (Specimens)

These lists include the identifiers of specimens that we strongly recommend are removed prior to any association testing. The simplest way to do this is to concatenate the files in to a single list and then to use PLINK's –remove command.

Exclusions_Specimens_Empirical_Sex_Test.txt
Exclusions_Specimens_IBS_PIHAT.txt
Exclusions_Specimens_missingness_and_f_score.txt
Exclusions_Specimens_nearest_neighbour_test.txt

Exclusion lists (Features)

These lists include the identifiers of features that we strongly recommend that you remove prior to analysis. This can be done by concatenating the files and then using PLINK's –exclude command. Depending on the application you may also want to filter your data set on the minor allele frequency (PLINK's –maf command) and proportion of missing data (PLINK's –geno command).

Exclusions_SNPs_Duplicated_features.txt
Exclusions_SNPs_snpflip_ambiguous.txt

IBS data

IBS_average_PIHAT.txt
This file contains the data on the average PIˆ score of each individual, as measured against every other individual

IBS.genome.genome.gz
This file contains the raw pairwise IBS analysis. Details of this file format can be found on the PLINK website.

SNPFLIP results

snpflip_output_initial.ambiguous snpflip_output_initial.annotated_bim
snpflip_output_initial.reverse
These files contain the results of the SNPFLIP analysis

Files supporting this document (Post-Imputation)

Raw data files

During the imputation, we handled the data in individual chromosomal sections. The raw data files are contained in gzipped tar files with self-explanatory names.

chr_1_imputed_data.tar.gz
chr_2_imputed_data.tar.gz
chr_3_imputed_data.tar.gz
chr_4_imputed_data.tar.gz
chr_5_imputed_data.tar.gz
chr_6_imputed_data.tar.gz
chr_7_imputed_data.tar.gz
chr_8_imputed_data.tar.gz
chr_9_imputed_data.tar.gz
chr_10_imputed_data.tar.gz
chr_11_imputed_data.tar.gz
chr_12_imputed_data.tar.gz
chr_13_imputed_data.tar.gz
chr_14_imputed_data.tar.gz
chr_15_imputed_data.tar.gz
chr_16_imputed_data.tar.gz
chr_17_imputed_data.tar.gz
chr_18_imputed_data.tar.gz
chr_19_imputed_data.tar.gz
chr_20_imputed_data.tar.gz
chr_21_imputed_data.tar.gz
chr_22_imputed_data.tar.gz

The files contained in each tar file are identical, albeit they differ in terms of the content.

Example data files in chr_22_imputed_data.tar.gz

22_impute2.bed
22_impute2.bim
22_impute2.fam
22_impute2.log

These files are the final output from the imputation analysis and are the ones that you will be most interested in. The files are derived from the 22.gen.gz file (see below) and are presented in standard PLINK format, ready for use in association testing against your favourite ELSA outcome variables. Please note that these files are of almost no value without also using the info file (described below) because around 80% of the data contained in them is low quality imputations that should be removed. For instance in chromosome 22, we imputed 1,111,880 SNPs, but just 20% of these (221,126) had an appreciable minor allele frequency and an info score above 0.8

Please note that it is up to you to remove the SNPs that you do not feel reach quality thresholds of your own determination. We provide the raw imputed data (for the benefit of advanced users).

22.info.gz
This file is absolutely critical to doing a robust analysis. It contains the quality data that you will definitely need to consider when using the imputed genotypes.

One of the most interesting thing about this file is the 'info' score, which is described on the IMPUTE2 website but which in simple terms you can interpret as a direct estimate of the proportion of imputed genotypes

at a locus that are likely to be correct. For instance if the info score for a SNP is 0.235 then you probably shouldn't include it in your analysis. We recommend filtering your data set to use only SNPs where the info score is a minimum of 0.8. For some users with limited computing power it may be sensible to carve the good SNPs out of the PLINK files using the PLINK –extract command, having first made a list of good SNPs using the info file. For people with a decent amount of RAM and a simple analysis (i.e. not permutation based or using complex mixed models), you might just as well perform the PLINK association testing on the whole lot, then bind/merge the PLINK .assoc file with the info file to get a lovely tidy final data set that contains all the scores on the coefficients, standard errors, P values, info, concordance on type 2 SNPs in a single file. It is then a simple process to drop lines of the data that don't meet the quality threshold.

22.gen.gz
This file contains the raw imputed genotypes but not in the format that many will have encountered before. See the IMPUTE2 documentation for an explanation of this file. It can be useful for people who are good at maths and are comfortable working with probabilities rather than 'hard calls'.

22.haps.gz
This file contains the estimated haplotypes and is the crucial file used by IMPUTE to estimate the probabilities for the imputation process.

22.summary.gz
This file could be useful if you wish to fine map a region or have some very exciting association data. It is a concatenated set of summary data from the IMPUTE2 algorithm. The most useful information in here is the data on the concordance test which validates the quality of the imputation process across the small regions of the chromsome that are imputed in each iteration of IMPUTE2. In lay terms this means that you should (just in case) check the quality of the imputation in specific regions that you are excited about before you start sending papers to Nature.

22.warnings.gz
Most of these warning files are empty. This is good. Any errors during imputation will be listed here.

22.bed
22.bim
22.fam
22.log
These four files are simply the chromosome 22 data prior to imputation. It is a simple slice through the original 001_elsa data set and this set of files is the input for the imputation.

22_impute2.nosex
22.nosex
22.sample
22.shapeit.ind.mm
22.shapeit.log
22.shapeit.snp.mm
These files are not much use to the majority of people. If you know what they are then you should also know what they are for. If not, please feel free to ignore them.

nohup.out
These files are artefacts of the script that was used to do the analysis. They don't make much sense and will only serve to confuse you. Do not try to use them.