GENEVA Genes and Blood Clotting Project
Quality Control Report

University of Washington

September 17, 2010

Contents

List of Figures

List of Tables

1 Summary and recommendations for dbGaP users

A total of 1,168 study subjects from the University of Michigan student community were genotyped on the Illumina Omni1-Quad v1_B array. The median missing call rate is 0.0395% and the mean error rate estimated from 36 pairs of sample duplicates and 14 pairs of monozygotic twins is $1.4e-05$.

Genotypic data are provided for all subjects (except two of questionable identity) and for all SNPs. However, we recommend selective filtering of genotypic data prior to analysis to remove sample-chromosome combinations with chromosomal anomalies and/or missing call rate $> 5\%$. The recommended filters are provided (Appendix A) along with filtered and unfiltered PLINK data files. Preliminary association test results are provided as an example of how to apply the filters. All SNPs are included in the association test results file, but we recommend that these be filtered according to the criteria specified in Table 1. A composite SNP filter is provided, along with each of the component criteria so that the user may vary thresholds. Additional specific recommendations are highlighted in the following document in *italics*.

2 Subjects and Phenotypes

The goal of this project is to identify genetic determinants of a variety of complex traits related to thrombosis, hemostasis and common human phenotypes through genome-wide association and genome-wide linkage analyses. Healthy subjects, ages 14 to 35, with at least one healthy full sibling, were recruited between June 26, 2006 to January 30, 2009 through targeted email to the University of Michigan, Ann Arbor, student population. The primary phenotype of interest is von Willebrand Factor antigen levels which were measured by the Alphalisa ™assay. Other traits were determined by a 52-question survey (bleeding traits and common traits) and the Bayer Advia 120 blood analyzer (peripheral blood traits).

3 Genotyping process

DNA samples were genotyped using the Illumina Human Omni 1M Quad v1_B SNP array and the BeadStudio calling algorithm at the Broad Institute Center for Genotyping and Analysis (CGA). DNA was extracted from buffy coat samples shipped to the Broad Institute Biological Samples Platform using the QiaCube methodology (QIAGen), then plated for Illumina production on 96-well plates. Plating of samples for production was not specifically randomized.

Production occurred within the Broad Institute's Genetic Analysis Platform. Each plate contained a HapMap CEU control placed in a random well. Genotypes were called from intensity data using Illumina BeadStudio (framework version 3.1.3.0) and genotyping module version 3.2.32. The Broad Institute considered a SNP as a technical failure if any of the following criteria pertained: call rate (over scans) $< 97\%$, number of replicate sample genotype discordances > 2, and number of sample trio inheritance errors > 1. Further, BeadStudio metrics GenTrain score (< 0.6) and cluster separation (< 0.4) values were also used as filters. All genotypes were converted to missing for technically failed SNPs.

4 Quality control process and participants

Genotypic data that passed initial quality control at the Broad were released to the GENEVA Coordinating Center (CC), the NCBI dbGaP team, and the Ginsburg study team. These data were further analyzed by these groups and discussed in weekly conference calls, which also included NHGRI personnel. Key participants in this process and their institutional affiliations are given in Appendix B.

Analysis tools varied by group, but include primarily PLINK [1] and the R statistical programming language [2]. If not otherwise noted, analyses described below were done using R and the "ncdf" library to access data stored in netCDF files (Section 19).

In the following QC summary, specific recommendations for dbGaP users are in *italics*.

5 Sample and participant number and composition

A total of 1,252 samples, including HapMap controls, were put into genotyping production, of which 1,220 were successfully genotyped and passed the Broad's QC process, as shown in Table 2. The subsequent QC process identified a further two scans for exclusion from dbGaP posting. Both of these samples had questionable identity that could not be resolved.

Genotyping scans are posted on dbGaP in two groups, (a) an analysis-ready set with just one scan per subject (and just one member of each pair of MZ twins), the one with lowest missing call rate, for both study subjects and HapMap controls and (b) duplicate scans, which include one member of each pair of MZ twins. The analysis-ready set consists of 1,155 scans from 1,152 different study participants and three different HapMap control subjects.

Among the study subjects, 36 are replicated twice, and there are 14 pairs of monozygotic twins, as shown in Table 3. The overall pedigree contains 2,190 subjects, which includes a number of subjects for which genotypes are not available, but phenotype data may be available. The majority of the ungenotyped samples in the pedigree are parents , so as to complete the sibling pedigrees. Among the 1,152 genotyped study subjects, there are 13 singletons, 366 families with two siblings each, and 124 families with three to six siblings each.

The 1,152 study participants are 62.59% female and 37.41% male. Sex was used as a covariate in the precompute association tests, as explained in Section 18.

6 Chromosomal anomalies

Gross chromosomal anomalies, such as aneuploidy and large insertion/deletion events, were detected by analyzing relative intensity ("LogRRatio") and a measure of allelic imbalance ("BAlleleFreq") [3]. BAlleleFreq is a transformation of the polar coordinate angle of the intensities of the two SNP alleles. Figure 1 shows that the occurrence of trisomic cells (or a mixture of monosomic and disomic cells) results in two different positions for heterozygotes at different loci. Peiffer et al. (2006) describe a transformation of the polar coordinate angle θ, which they call "BAlleleFreq" (BAF). This value is one of the metrics output by the Illumina BeadStudio software. Figure 2 shows that BAF is a transformation that standardizes the positions of the three diploid genotypic classes to 0 (AA), 0.5 (AB) and 1.0 (BB).

Figure 3 shows normal BAF and LRR scans for chromosome one in sample A, where each point represents a different locus in one sample. The plot in Figure 3b has bands at approximately 0, 0.5 and 1.0. The bands at 0 and 1.0 consist of loci in which this sample is homozygous and the band at 0.5 is the BAF value for loci at which sample A is heterozygous.

To identify aneuploid or mosaic samples systematically, we calculated, for each sample, the variance of the BAF values for heterozygous SNPs in a sliding window along the genome. Each chromosome was divided into bins approximately 8Mb wide and each window consists of two consecutive bins. The window slides along in one-bin increments. We then examined chromosome scans (of the type shown in Figure 3) for all sample-chromosome combinations that have one or more windows for which the heterozygous BAF variance is greater than four standard deviations from the mean of all variances for that window. We also examined samples that are outliers in mean probe intensity on each chromosome relative to other chromosomes for the same sample (i.e. more than 6 interquartile ranges from the upper or lower quartiles). No autosomal anomalies greater than 10 Mb were found. However, one sex chromosome anomaly was detected, as described in Section 7.

We recommend filtering out the genotypes for all SNPs in sample-chromosome combinations with a chromosomal anomaly. The identifiers are provided in the file "chromosome.anomalies.csv" on dbGaP. See also Appendix A.

In addition to flagging samples detected by a split in the intermediate, heterozygous band, we also examine BAF/LRR plots for samples for which all chromosomes are flagged. This could indicate a contaminated sample that consists of DNA from more than one participant, which typically has more than the expected three bands in BAF plots for all chromosomes. We also examined BAF/LRR plots for samples that are outliers for either high or low heterozygosity. None of these samples had abnormal patterns or appeared to be contaminated.

7 Gender identity

To verify that the annotated gender from the study site database and the genetic gender are consistent, we look at the mean of the intensities of the SNP probes on the X and the Y chromosomes, along with the X chromosome heterozygosity, as shown in Figure 4. The expectation is that males and females will group into two distinct clusters that differ markedly in both the X and Y intensity values. Figure 4d shows the results for all samples prior to exclusions based upon questionable identity. There are two clusters, as expected, and 12 samples that appear to have a gender discrepancy, i.e. the sample lies in the 'wrong' cluster based upon the annotated gender. All but one of the gender discrepancies was resolved in conjunction with the relatedness analysis, as described in Section 8. The remaining subject with a gender discrepancy and their full sibling were excluded from the data set posted on dbGaP.

The X versus Y chromosome intensity plot in Figure 4a shows one male sample with unusually high Y chromosome intensity. Examination of the BAF profile (Figure 5) of the pseudo-autosomal (XY) chromosome probes in this sample shows a split intermediate band, which is evidence of allelic imbalance expected for three copies of the Y chromosome, suggesting an XYY or XY/XYY mosaic.

8 Relatedness

The relatedness between each pair of participants was evaluated by estimation of three coefficients corresponding to the probability that zero (*k0*), one (*k1*) or two (*k2*) pairs of alleles are identical by descent (IBD). The kinship coefficient (KC) for a pair of participants is

$$KC = \frac{1}{2}k2 + \frac{1}{4}k1 \qquad (1)$$

Table 4 shows the expected coefficients for some common relationships. Any two alleles at a locus are either identical by descent or not and this gives rise to variation of actual identity around the expected values. When markers over the entire genome are used to estimate the kinship coefficient, there is a need to take into account the dependencies among markers due to linkage. Expressions for the variance in a summary measure of actual identity have been given in the past [4, 5, 6, 7] and have been extended by the CC to the three IBD coefficients. The expected values ± 2 standard deviations are given in Figure 6 for full siblings, half siblings and first cousins and are indicated with orange bars.

For this study, the IBD coefficients were estimated using 115,045 autosomal SNPs and the method of moments procedure used by PLINK [1], but implemented in R. The SNPs were selected at random from all autosomal SNPs with missing call rate less than 5% and minor allele frequency > 0, with the constraint that no two SNPs are closer than 15 kb apart.

Figure 6 is a plot of the estimated IBD coefficients, *k0* and *k1*, color-coded by inferred relationship. This plot shows all pairs of subjects with KC$> 1/32$ (half the expected value for first cousins). The inferred relationships were consistent with the original pedigree annotations except for 28 pairs of unexpected duplicates, 40 pairs expected to be full-siblings but appearing unrelated (or vice-versa) and three pairs expected to be full siblings, but appearing as half-sibs. All of these discrepancies (except the half-sibs) were resolved by examining the pattern of unexpected relationships and consulting the records. In principle, the three pairs of samples near *k0*$= 0.5$ and *k1*$= 0.5$ could be half-sibling, avuncular or grandparent-grandchild relationships. However, we corrected the pedigree assuming they are half-siblings for three reasons: (a) similiarity in age, (b) the study specifically recruited siblings and (c) sharing of mitochondrial SNP genotypes indicated that the two members of each pair have the same mother.

For an analysis that assumes all participants are unrelated, we recommend selecting one participant from each family-unit. See Appendix A.

9 Population structure

To investigate population structure, we use principal components analysis (PCA), essentially as described by Patterson et al [8]. PCA is problematic in studies with a large number of related pairs of subjects. One possible solution to this problem is described by Zhu et al [9]. They suggested calculating both sample and SNP eigenvectors for a subset of unrelated subjects and then using the resulting SNP eigenvectors, along with the genotype calls for the remaining subjects, to estimate their sample eigenvectors. In essence, we construct SNP eigenvectors by using only unrelated subjects, then project all subjects along these fixed SNP eigenvectors to obtain sample eigenvectors. We have implemented this approach for this study of sibships. In the following, we use the term “direct estimation” to refer to the usual method of calculating sample eigenvectors directly from the genotype calls alone and we use the term “indirect estimation” to refer to calculating sample eigenvectors from genotype calls and SNP eigenvectors that were calculated from related subjects.

PCA was performed on four different sample sets, summarized in Table 5. We used two sets of SNPs, one defined using minor allele frequency on all study subjects and one defined using minor allele frequency for the self-identified White samples. To select SNPs for the first set, we started from a pool of 722,810 autosomal SNPs with missing call rate $< 5\%$ and minor allele frequency $> 5\%$, as calculated from all study samples. From this pool, we selected 111,602 SNPs in two rounds of linkage disequilibrium (LD) pruning using the pair-wise genotypic correlation method in PLINK [1]. In the first round, our goal is to remove short-range LD, thus, we use a window size of 50 SNPs with a five SNP offset and an r^2 threshold of 0.2. The second round of pruning is intended to remove long-range LD. Historically, we and others [10, 11] have found that long-range LD can produce eigenvectors that are highly correlated with local clusters of SNPs. Therefore, the second round of pruning uses a window size equal to the median number of SNPs in 5Mb from the first round. In this case, the median value was 434 SNPs with an offset of 5Mb. The r^2 threshold remained at 0.2 for the second round of pruning and the result was 111,602 SNPs to use for PCA. For the second SNP set, we started with a pool of 708,444 autosomal SNPs that satisfied the same criteria, but with an allele freqency > 0.05 as calculated from self-identified White samples. After two rounds of LD pruning using PLINK, the result was 103,510 SNPs to use for PCA.

In some PCA analyses, we included 209 additional HapMap founder samples external to the HapMap control samples genotyped with the study subjects. This allowed detection of population group outliers among the study subjects using the known HapMap population anchor points. Data for the external HapMap samples genotyped on the Illumina Omni 1M Quad platform were obtained from the Illumina web site (see Section 19).

As an additional SNP filter, we calculated sample discordance per SNP in HapMap samples, comparing the genotype data we received with the data posted by Illumina. There is a total of 16 scans from three HapMap samples that were genotyped in this study and in the external HapMap set. For SNP set one, there were 56 probes with > 0 discordant calls, and 46 probes with > 0 discordant calls in SNP set two. Thus, our final SNP set counts are 111,546 and 103,464, respectively.

Initially, we analyzed 502 unrelated study samples along with the 209 HapMap founders in order to identify population group outliers (PCA set 1). A plot of the first two eigenvectors from this analysis is shown in Figure 7. A majority of samples are self-identified as White and they cluster with the CEU subjects, as expected. There are small numbers of many other self-identified ethnicities among the study samples that diverge from the cluster of self-identified Whites. For example, self-identified Black subjects form a trail of points extending from the African HapMap controls towards the CEU and White study subjects. The Asian Indian subjects form a trail extending from the White study subjects towards the CHB and JPT HapMap controls. The placement of samples in this PCA plot generally agrees with self-reported ethnicity.

Next, we performed PCA on the same set of 502 unrelated study samples, but without HapMap samples (PCA set 2). We used the same set of 111,546 LD-pruned SNPs. A plot for the first two eigenvectors is shown in Figure 8 along with median and half-standard deviation lines for self-identified White subjects. From the results we defined a homogeneous set of samples with PCA-defined European ancestry. Samples that fell outside one-half standard deviation (SD) from the median for eigenvector one and one-half SD from the median for eigenvector two were excluded. This resulted in a set of 406 unrelated study samples of PCA-defined European ancestry. This sample set was used for the Hardy-Weinberg Equilibrium (HWE) tests as explained in Section 14. We ran HWE with this sample set in order to reduce or eliminate population structure so that the HWE test will detect mainly genotyping artifacts, because the HWE test can be affected by population structure and by inclusion of related individuals.

In order to obtain eigenvectors for all study subjects (related and unrelated), we implemented the method described by Zhu et al [9]. Before applying the method to the study data as desired, we first conducted an exercise to validate our implementation of the method. We took two members from each of the study families and randomly assigned one member to set A and one member to set B. We calculated PCA results directly for set A (as described above), then used the Zhu method to indirectly calculate sample eigenvectors for the members of set B. Then, we did the reverse: calculated PCA results directly for set B and indirectly for set A. Plots comparing the first two direct and indirect eigenvectors for set A and set B samples are shown in Figures 9 and 10. Both eigenvectors one and two show a high correlation between direct and indirect calculation, indicating that the inferred eigenvectors for related samples will be useful when adjusting for population structure in association tests. The r^2 for set A is 0.9929 for eigenvector one and 0.7879 for eigenvector two. For set B, the r^2 is 0.9937 and 0.7959 for eigenvectors one and two, respectively.

For both the set of 502 unrelated study samples and the set of 406 unrelated, PCA-defined European samples, we estimated the sample eigenvectors for the remaining related samples from the study. To do this we used the directly calculated results to infer the eigenvectors for the remaining family members. Figures 11 and 12 show the combined direct and inferred results for all samples from the analyses of all (set 3) and of European-ancestry subjects (set 4), respectively.

We used the fourth eigenvector as part of the adjustment for the preliminary association test, for this was the most significant of the eigenvectors to be associated with the primary outcome, von Willebrand Factor (VWF) level.

The first 20 eigenvectors from PCA sets three and four are provided as sample-by-eigenvector matrices on dbGaP and labeled as "Principal_components_study.csv" and "Principal_components_Euro_study.csv". The sample sets are identified by the row labels in these files, as well as variables in the sample analysis results table ("Sample_analysis_results.csv") with names "pca.rel.study" (PCA set 3) and "pca.euro" (PCA set 4). The SNP sets are identified by the variables in the SNP table ("SNP_analysis_results.csv") with variable names "pca.study.set" and "pca.euro.set".

10 Missing call rates

Two missing call rates were calculated for each sample and for each SNP and are provided in files "SNP_analysis_results.csv" and "Sample_analysis_results.csv" on dbGaP. *missing.n1* is the missing call rate per SNP over all samples including the HapMap controls. *missing.e1* is the missing call rate per sample for all SNPs with *missing.n1*$<$ 100%. *missing.n2* is the missing call rate per SNP over all samples with *missing.e1*$<$ 5%. In this project, there were no samples with *missing.e1*$>$ 5%, thus the values of *missing.n1* and *missing.n2* are identical. Finally, *missing.e2* is the missing call rate per sample over all SNPs with *missing.n2*$<$ 5%.

Figure 13 shows the distribution of *missing.e2* for all study samples. The median value is 0.0395%, the 95^{th} percentile is 1.6958% and the maximum *missing.e2* value is 2.8420%.

The Illumina Omni 1M Quad array contains 1,016,423 assays for SNPs. Prior to data release, the Broad QC process failed 36,568 SNPs with call rate $< 97\%$ and other criteria (Table 1). Among the remaining 979,855 unfailed SNPs, the median missing call rate is 0.0000% and the 95^{th} percentile is 0.00003%.

11 Batch effects

To look for genotyping batch effects on the missing call rate per sample, we analyzed *missing.e1* for autosomal SNPs. In this project, each genotyping batch is a group of samples that occupy a 96-well plate and are processed together through the genotyping chemistry, resulting in up to 96 samples per group. For simplicity, we will refer to a group of samples processed together as a 'batch'. Each plate generally contains 95 study samples and all of the batches contain one CEU HapMap control. Sample failures, reruns and remnant sample runs can result in plates with fewer than 95 study samples.

The study samples were processed in 14 batches. After removing failed samples, the median number of samples per batch is 95, and the range is from 41 to 96, including HapMap controls. Although the batch effect on missing call rate is significant, all plates have low missing call rates. The median of missing call rates for each of the 14 batches ranges from 0.0304% to 0.0530%. Figure 14 shows a plot of the number of samples per batch versus the median missing call rate per batch, for autosomal SNPs.

Another way to test for batch effects is to assess the difference in allelic frequency between each batch and a pool of all other batches [12]. We calculated a one degree of freedom χ^2 test statistic for each SNP and batch and then averaged over SNPs within a batch. This statistic is a measure of how different each batch is from the other batches. It can be affected not only by laboratory processing but also by the biological composition of samples within a batch. The characteristic most likely to affect the distinctiveness of a batch is the ethnic composition relative to the mean composition across batches. Figure 15 shows a plot of genotyping plate composition versus the mean χ^2 statistic, where plate composition is measured as the fraction of samples in the plate that are self-identified as White. No plates appear to be problematic with regard to the allelic frequency test.

12 Tests for phenotypic associations with missing call rate

Linear regression was used to test whether the log_{10} of the autosomal missing call rate is significantly associated with varying levels of VWF. There is no significant association (p-value=0.702).

13 Duplicate discordance

A very useful measure of genotyping accuracy is the duplicate discordance rate. A total of 36 study samples were genotyped in duplicate. Further, there are 14 pairs of monozygotic twins from which we are also able to use to calculate duplicate discordance. Figure 16 shows the distribution of discordance rates with the monozygotic twin pairs colored in magenta; the median discordance is 0.00291% and the range is 0.00041% to 0.04041%.

Genotyping error rates can be estimated from duplicate discordance rates. The genotype at any SNP may be called correctly, or miscalled as either of the other two genotypes. If α and β are the error rates associated with the two miscalled genotypes, the probability that duplicate genotyping instances of the same participant will give a discordant genotype is $2[(1-\alpha-\beta)(\alpha+\beta)+\alpha\beta]$. When α and β are very small, this is approximately $2(\alpha+\beta)$ or twice the total error rate. Potentially, each true genotype has different error rates (i.e. three α and three β parameters), but here we assume they are the same. In this case, since the overall discordance rate is about 0.000029, a rough estimate of the mean error rate is $1.4e-05$.

Duplicate discordance estimates for individual SNPs can be used as a SNP quality filter. The challenge is to find a level of discordance that would eliminate a large number of SNPs with high error rates while retaining a large number of SNPs with low error rates. *We recommend a filter threshold of* > 0 *discordant calls because this retains* $> 90\%$ *of SNPs with an error rate* < 0.001*, while removing over* 63% *of SNPs with an error rate of* 0.01. Refer to Table 6.

14 Hardy-Weinberg equilibrium

We calculated an exact test of Hardy-Weinberg equilibrium (HWE) using a set of 406 unrelated study samples with primarily European ancestry as defined by PCA, all with missing call rate < 0.02. European ancestry subjects were defined as those within one-half a standard deviation (SD) of the median of eigenvector one and within one-half SD from the median value of eigenvector two as found in the PCA analysis with all unrelated study subjects, described in Table 5 and Figure 8. The samples listed in "hwe.euro.keep.txt" were used for this analysis in conjunction with sample-chromosome filters outlined in "jcmat.anom.miss.05.csv". The p-values were calculated for each SNP and compared to the expected value in the QQ plots shown in Figure 17 for autosomal and X chromosome SNPs separately. At approximately 1e-3, the p-values for the autosomal SNPs become inflated, however there is no substantial deviation in the X chromosome SNPs. Rather, a handful of probes have extreme values, all of which have all samples called as heterozygote.

Population structure usually results in a deficiency of heterozygotes and a positive inbreeding coefficient. We estimated the inbreeding coefficient as

$$1 - \frac{\text{number of observed heterozygotes}}{\text{number of expected heterozygotes}} \tag{2}$$

The mean of this estimate for autosomal SNPs for the above described European sample set is 0.00104. The distribution is roughly symmetrical around 0, as seen in the plots in Figure 18, suggesting that there is little population structure. We conclude that most deviations from Hardy-Weinberg may be due to genotyping artifacts.

Although the QQ plots in Figure 17 show deviation of observed from expected p-values for autosomal SNPs a bit below 0.001, *we suggest using a filter threshold of p=0.0001 because examination of cluster plots reveals good cluster plots for many assays with p-values* > 0.0001. The threshold is rather subjective, but we are reluctant to recommend a higher threshold that would eliminate many good SNP assays.

15 Sample exclusion and filtering

As mentioned previously, genotyping was attempted for a total of 1,236 samples, of which 1,204 passed the Broad QC process. The subsequent data cleaning QA process identified 2 subject scans that will not be included in the dbGaP posting, both of which had sample identity issues explained in Section 7 and Section 8. The remaining 1,202 study scans will be posted on dbGaP with all accompanying files.

For association testing, we recommend filtering out SNPs for each sample-by-chromosome combination with an autosomal chromosomal anomaly (as described in Section 6) and/or with a missing call rate per chromosome $\geq$ 5%. *Also, it is recommended to filter out SNPs on all chromosomes for each sample with a sex chromosome anomaly as well as samples with an overall missing call rate* $\geq$ 2%. Table 7 shows that the percentage of potential genotype calls lost over the autosomes is 0.000%, 0.173% for the X chromosome, 0.463% for males only on the Y chromosome and 1.558% on the mitochondrial genome.

For specific analyses, such as Hardy-Weinberg testing, additional filters are suggested, such as filtering out all samples except one sample per participant, one participant per family, and one ethnic group. To facilitate the application of these filter suggestions, we provide a sample-by-chromosome matrix to remove specific chromosomes from specific samples, as well as whole sample filter vectors. Most of these are logical, where TRUE indicates retention and FALSE indicates exclusion. We also provide a PLINK analysis-ready data set with the sample-chromosome filters applied and “keep” files for applying whole sample filters. See Appendix A.

16 Mendelian errors

When multiple parent-offspring trios or pairs have been genotyped, Mendelian errors provide another useful SNP quality metric. However, this study has only one HapMap control trio, so Mendelian errors were not used, since we have observed that chromosomal aberrations specific to a single trio can cause multiple Mendelian errors.

17 SNP filters

Table 1 summarizes a sequence of SNP failure criteria applied by the Broad QC process prior to data release and a set of additional filters suggested for removing assays of low quality or informativeness. *The suggested composite quality filter is provided as a logical vector in the files of association test results discussed in Section 18*. These filters remove 5.53% (85,196) of the 1,016,423 SNP assays attempted. Figure 19 shows the overlap of SNPs filtered out when using four of the recommended filters. There is relatively little overlap, indicating that each these filters are detecting different types of assay problems. In addition to the quality filters, *we also suggest applying a minor allele frequency filter* for viewing association test results. The frequency criterion depends upon power considerations, but is usually on the order of 1-5%. In Table 1 we show the effect of a 2% MAF filter on each of the sample sets used for the preliminary association tests. *Regardless of what filters are applied to association test results, it is highly recommended to view SNP cluster plots for any SNPs of interest.*

18 Preliminary association test results

For this study of sibships, we used PLINK's QFAM association testing program to account for relatedness in a set of preliminary association tests. We also used a linear regression model, ignoring relatedness, for comparison and to provide a set of results that can be replicated easily by dbGaP users. The primary trait of interest is VWF level (log_{10} transformed to achieve normality). We tested for an association between $log_{10}(VWF)$ and the potential covariates: sex, age and eigenvectors 1-32. Only sex and eigenvector four (EV4) were significant. Because QFAM does not allow covariates, we adjusted $log_{10}(VWF)$ for sex and EV4 prior to both analyses. QFAM uses permutations to account for the dependence between related individuals. Initially we performed 10^5 permutations and subsequently selected 93 SNPs with p-values less than 10^{-4} for additional permutations. The analyses performed are summarized in Table 8.

For model A1, all autosomal SNPs were tested with 100,000 permutations, for which the smallest possible p-value is 10^{-5}. The QQ plot for this test is shown in Figure 20. The plot is truncated at 5 on the y-axis, as expected, because of the value limit based upon the number of permutations. Figure 21 shows the Manhattan plot for this model. The p-values shown are for SNPs filtered for quality as described in Table 1.

There are 93 SNPs with a p-value $< 10^{-4}$. We ran PLINK QFAM again with these 93 probes using 10^8 permutations, because as the permutations increase, the p-values for these SNPs will get increasingly close to their true value. A comparison of the p-values from 100,000 permutations and 10^8 permutations are shown in Figure 22. One can see that most, but not all, p-values get smaller as the number of permutations increases. A Manhattan plot created with the updated p-values for the 93 top probes is shown in Figure 23. The p-values for the 93 SNPs are stored in “Association_test_top93_PermutationResults.csv”.

A linear regression model was also used to obtain preliminary association test results, ignoring relatedness. The same sample set and adjusted outcome variable were used. Each SNP effect was analyzed with the genotype coded as 0, 1 and 2 (additive genotypic model). SNP effects were assessed with likelihood ratio tests. Figure 24 shows that the observed p-values are inflated, with a genomic control factor of 1.158 (compared to a genomic control factor of 0.9896 for the QFAM analysis). This inflation is expected from ignoring relatedness. Further, the Manhattan plot, Figure 25, shows more significant p-values than seen with the PLINK QFAM results.

Cluster plots of the top 27 hits from all models, which are the same 27 SNPs, are provided in “top_hits_cluster.pdf”. These plots show very good clustering of the most significant SNPs.

19 URLs

netCDF data files: http://www.unidata.ucar.edu/software/netcdf/
Illumina data files: http://www.illumina.com/

References

[1] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, and M.A. Ferreira et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81:559–575, 2007.

[2] R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria*, ISBN 3-900051-07-0:URL http://www.R–project.org, 2006.

[3] D.A. Peiffer, J.M. Le, F.J. Steemers, W. Chang, and T. Jenniges et al. High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research*, 16:1136–1148, 2006.

[4] C.C. Cockerham and B.S. Weir. Variance of actual inbreeding. *Theoretical Population Biology*, 23:85–109, 1983.

[5] S.W. Guo. Variation in genetic identity among relatives. *Human Heredity*, 46:61–70, 1996.

[6] W.G. Hill. Variation in genetic identity with kinships. *Heredity*, 71:652–653, 1993.

[7] P.M. Visscher. Whole genome approaches to quantitative genetics. *Genetica*, 136:351–358, 2009.

[8] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2:e190, 2006.

[9] Xiaofeng Zhu, S. Li, R. S. Cooper, and R. C. Elston. A unified association analysis approach for family and unrelated samples correcting for stratification. *American Journal of Human Genetics*, 82:352–365, 2008.

[10] J. Novembre, T. Johnson, K. Bryc, Z. Kutalik, and A.R. Boyko et al. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.

[11] C. Tian, R.M. Plenge, M. Ransom, A. Lee, and P. Villoslada et al. Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genetics*, 4:e4, 2008.

[12] Anna Pluzhnikov, J. E. Below, A. Konkashbaev, A. Tikhomirov, E. Kistner-Griffin, C. A. Roe, D. L. Nicolae, and Nancy J. Cox. Spoiling the whole bunch: Quality control aimed at preserving the integrity of high-throughput genotyping. *The American Journal of Human Genetics*, 87(1):123 – 128, 2010.

A Sample filter vectors and sample-by-chromosome filter matrix

For association testing, we recommend filtering genotypic data using a sample-chromosome filter matrix and whole-sample filters.

The file “jcmat.anom.miss.05.csv” is a sample-by-chromosome matrix with 1,155 samples and 26 chromosomes (1-22 autosomes, X, pseudo-autosomal, Y and mitochondrion). Each element of the matrix is TRUE/FALSE to indicate retaining/eliminating each sample-by-chromosome combination for analysis. The elements of this matrix are FALSE when either the sample-chromosome combination has a chromosome anomaly as listed in “chromosome.anomalies.csv” and/or when it has a missing call rate > 0.05. PLINK-formatted genotypic data files are provided before and after the application of this filter matrix. *We recommend using the filtered version for all analyses (labeled as “zeroed-out”). In addition, we recommend whole-sample filters for each type of analysis.* The whole sample filters are described below and are given either in the file “Sample_annotation.csv” or “Sample_analysis_results.csv”, which have one sample per subject, including study subjects and HapMap controls.

1. *hwe.unrel.euro* indicates unrelated study samples of homogeneous European ancestry used in HWE (“Sample_analysis_results.csv” file)

2. *assoc_model1* indicates study samples used in the association test with non-NA value of *lvwf* (“Sample_analysis_results.csv” file)

3. *lvwf* is a vector of numeric values of the $log_{10}(VWF)$ levels for subject (“Sample_analysis_results.csv” file)

For the HWE and association tests, “keep” files are provided and were created using the above explained whole-sample vectors as described below.

1. “hwe.euro.keep.txt” lists samples for which *hwe.unrel.euro* is TRUE and *missing.e2* < 0.02. These are the samples used for Hardy-Weinberg equilibrium testing, Section 14.

2. “assoc.keep.txt” lists samples for which *assoc_model1* is TRUE and *lvwf* is non-NA. These samples were used for association testing as described in Section 18.

B Project participants

University of Michigan
Karl Desch, David Ginsburg, Jun Li

Broad Center for Genotyping and Analysis
Andrew Crenshaw, Daniel Mirel

dbGaP, NCBI
Mike Feolo, Justin Paschall, Nataliya Sharopova

GENEVA program, NHGRI
Teri Manolio

Department of Biostatistics, University of Washington
David Crosslin, Cathy Laurie, Cecelia Laurie, David Levine, Thomas Lumley, Caitlin McHugh, Sarah Nelson, Ken Rice, Jess Shen, Bruce Weir, Xiuwen Zheng

SNPs kept	SNPs lost	filter
1,016,423	0	SNP assays attempted
979,854	36,569	Broad: technical filters
934,996	44,858	CC: MAF=0 for all samples
934,137	859	CC: call rate $< 98\%$
933,745	392	CC: HWE p-value $< 10^{-4}$
931,235	2,510	CC: > 0 discordant call in 36 dup pairs & 14 pairs of monozygotic twins
931,233	2	CC: difference in autosomal allele frequency between sexes > 0.1
931,227	6	CC: sex difference in heterozygosity ≥ 0.3
94.47%	5.53%	Percentage of SNP assays attempted
779,702	236,721	CC: MAF < 0.02 in study participants used for association test
76.71%	23.29%	Percentage of SNP assays attempted for linear regression analysis

Table 1: Summary of recommended SNP filters. "Broad" refers to SNPs failed by the genotyping center and "CC" refers to filters recommended by the GENEVA Coordinating Center. Broad technical filters include: call rate $< 97\%$, GenTrain score < 0.6, cluster separation < 0.4, number of replicate errors > 2 and number of parent-parent-child (P-P-C) errors > 1.

Study	HapMap	
1,236	16	DNA samples into genotyping production
-32	-	Failed genotyping
1,204	16	Scans released by genotyping center
-2	-	Sample identity issues
1,202	16	Scans to post on dbGaP
1,202	16	Filtered scans for analysis

Table 2: Summary of DNA samples and scans.

Study Samples	HapMap	
1,202	16	Scans
1,152	3	Participants
36	13	Replicated Participants
14	-	Monozygotic Twin Pairs
503	1	Families

Table 3: Filtered genotype scan and participant numbers.

$k2$	$k1$	$k0$	Kinship	Relationship
1.00	0.00	0.00	0.5	MZ twin or duplicate
0.00	1.00	0.00	0.25	parent-offspring
0.25	0.50	0.25	0.25	full siblings
0.00	0.50	0.50	0.125	half siblings
0.00	0.25	0.75	0.0625	cousins
0.00	0.00	1.00	0.00	unrelated

Table 4: Expected identity-by-descent coefficients for some common relationships.

	Study Participants			
Set	Unrelated	Related	HapMap Founders	Total
1	502	-	209	711
2	502	-	-	502
3	502	650	-	1,152
4	406	536	-	942

Table 5: Sample sets used for PCA analysis. Unrelated sample eigenvectors were estimated using the “direct” method and related samples were estimated using the “indirect” method, as outlined in Section 9. Set 4 consists of all samples with PCA-defined European ancestry. Set 2 was used for the definition of samples for the Hardy-Weinberg equilibrium test as described in Section 14. All sets included only unduplicated samples.

	Assumed Error Rate				
# discordant calls	1e-5	1e-4	1e-3	1e-2	# SNPs
dis>0	0.00099950	0.00995041	0.09518519	0.63303283	2630
dis> 1	0.00000049	0.00004868	0.00459051	0.26144183	46
dis> 2	0.00000000	0.00000016	0.00014582	0.07706802	2
dis> 3	0.00000000	0.00000000	0.00000341	0.01732500	1

Table 6: Probability of observing more than the given number of discordant calls in 36 pairs of duplicate samples and 14 pairs of monozygotic twins, given an assumed error rate. The number of SNPs with a given number of discordant calls is shown in the final column. The row in red is the recommended threshold for SNP filtering, which is > 0 discordant calls.

Chromosome type	Chromosome anomalies	Missing call rate per chromosome $\geq 5\%$	Both	Sample Size (Number of SNPs)
Autosomes	0.000000%	0.000000%	0.000000%	953,725
X	0.000000%	0.173160%	0.173160%	23,934
Y	0.086580%	0.231481%	0.462963%	1,170
M	0.000000%	1.558442%	1.558442%	26

Table 7: Summary of recommended sample-chromosome filters. The percentage of genotype calls lost to each filter type is given.

Analysis	Sample Size	Model	Test Type	SNP set	Genomic Control Factor
A1	1,151	$residuals(log_{10}VWF \sim sex + EV4) \sim genotype$	PLINK QFAM, 1e5 permutations	autosomal SNPs	0.990
A2	1,151	$residuals(log_{10}VWF \sim sex + EV4) \sim genotype$	PLINK QFAM, 1e8 permutations	top 93 SNPs	-
A3	1,151	$residuals(log_{10}VWF \sim sex + EV4) \sim genotype$	linear regression	autosomal SNPs	1.158

Table 8: Regression models analyzed for preliminary association tests. Regression was done to assess the significance of any of the following on the log_{10}(VWF): sex, age, EV1-32. For those that were significant, we performed regression again with only those covariates and used the residuals from the model as the phenotype. The $EV4$ values are the fourth principal component as calculated from PCA sample set 3; see Table 5 for the PCA sample set definitions.

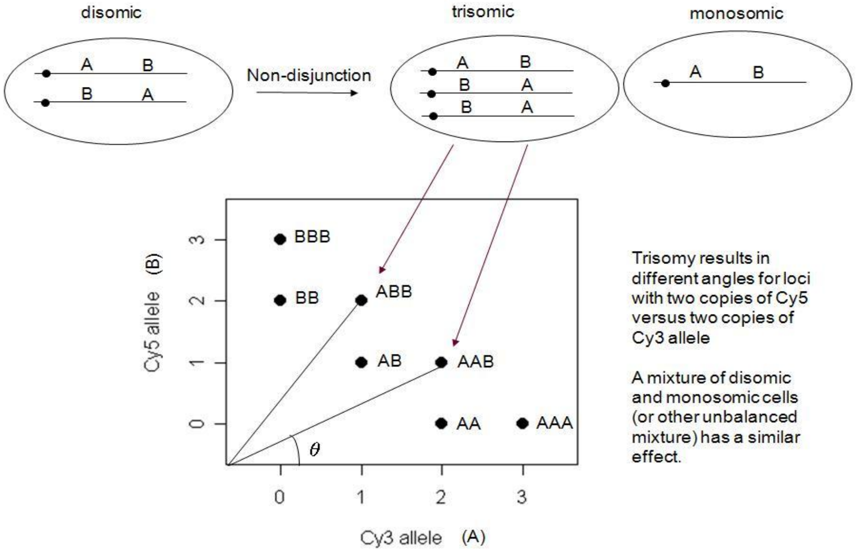
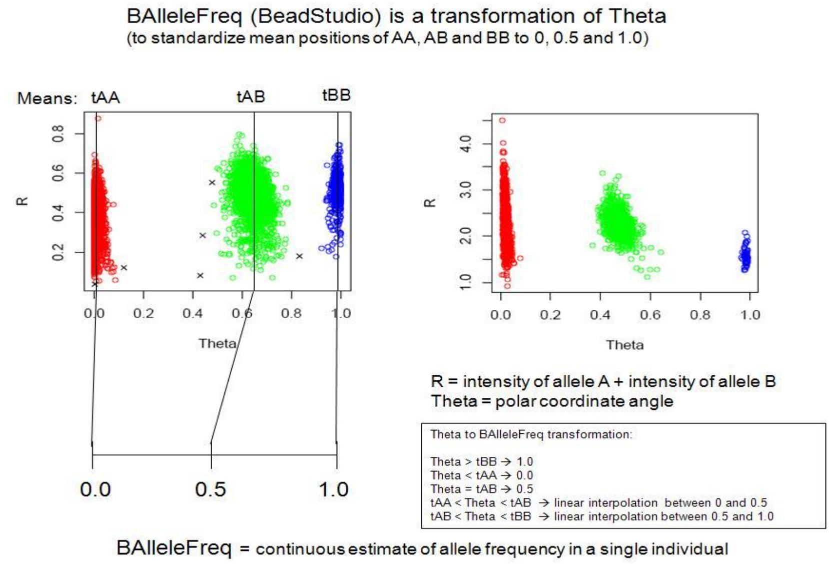

disomic
A
B
B
A
Non-disjunction
trisomic
A
B
B
A
B
A
monosomic
A
B
BBB
BB
ABB
AB
AAB
AA
AAA
θ
0
1
2
3
Cy5 allele (B)
0
1
2
3
Cy3 allele (A)
Trisomy results in different angles for loci with two copies of Cy5 versus two copies of Cy3 allele
A mixture of disomic and monosomic cells (or other unbalanced mixture) has a similar effect.

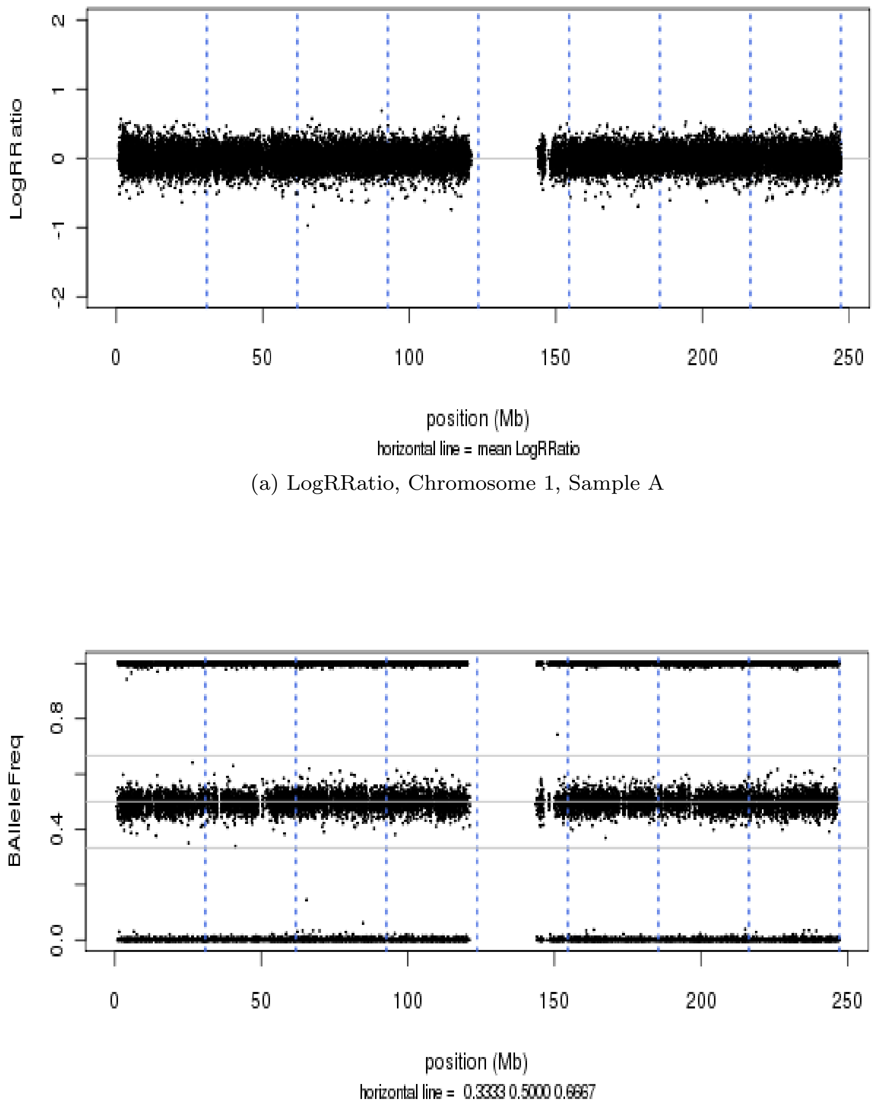
Figure 1: Schematic diagram showing how trisomic cells (or a mixture of disomic and monosomic cells) can results in two different polar coordinate angle positions for heterozygotes at different loci. “A” and “B” represent two alleles at one locus, where the former is tagged with Cy3 and the latter with Cy5. Loci with allele B on the duplicated chromosome have ABB heterozygotes and those with allele A on the duplicated chromosome have AAB heterozygotes.



Theta to BAlleleFreq transformation:

Theta > tBB → 1.0
Theta < tAA → 0.0
Theta = tAB → 0.5
tAA < Theta < tAB → linear interpolation between 0 and 0.5
tAB < Theta < tBB → linear interpolation between 0.5 and 1.0

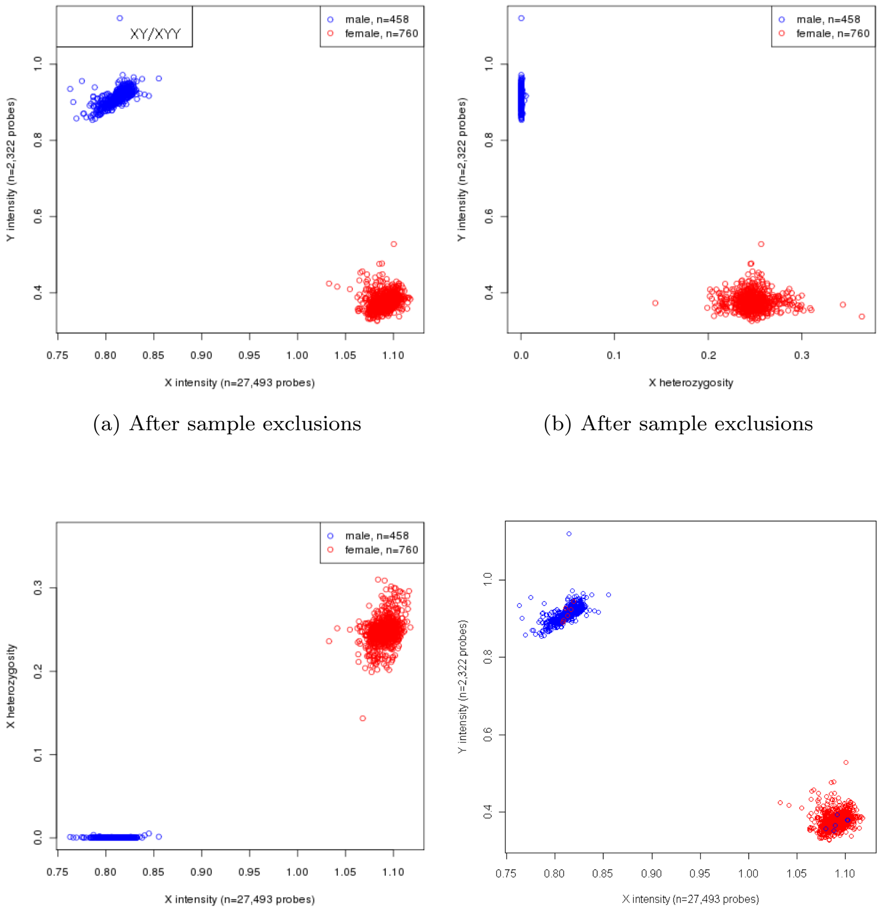
Figure 2: "BAlleleFreq" (BAF) is a transformation of the polar coordinate angle (θ) to standardize the positions of the three diploid genotypes to 0, 0.5 and 1. The quantities tAA, tAB and tBB are the mean θ values for each genotypic class. Values of θ less than tAA are assigned a BAF value of 0 for tAA, while those greater than tBB are truncated to a BAF value of 1. Values between tAA and tAB are positioned by linear interpolation between 0 and 0.5, and similarly for those between tAB and tBB. In the plot, red indicates an AA genotype, green implies a heterozygote locus and blue is a BB genotype. A black 'x' indicates a missing genotype call.



(a) LogRRatio, Chromosome 1, Sample A

(b) BAlleleFreq, Chromosome 1, Sample A

Figure 3: Figure 3a shows a normal LogRRatio scan of chromosome 1 for sample A. The values have a mean roughly around zero across the entire chromosome. Figure 3b shows a normal BAlleleFreq scan of for the same sample and chromosome. Each point represents a SNP. The upper and lower bands represent homozygotes and the intermediate band represents heterozygotes.



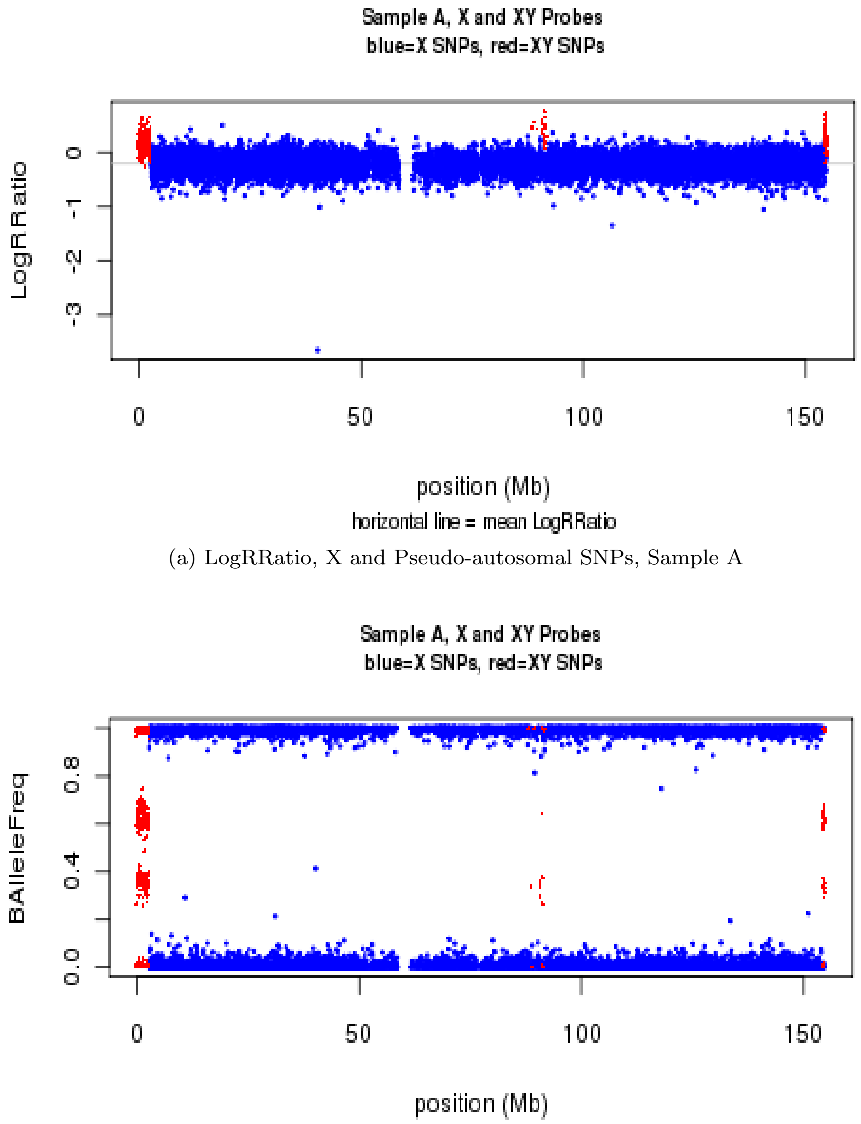
(a) After sample exclusions

(b) After sample exclusions

(c) After sample exclusions

(d) Before sample exclusions

Figure 4: The X and Y intensities are calculated for each sample from the mean of the sum of the normalized intensities of the two alleles for each probe on those chromosomes. Sample sizes are 27,493 for the X chromosome and 2,322 for the Y chromosome. X heterozygosity is the fraction of heterozygote calls out of all non-missing genotype calls on the X chromosome for each sample. Inferred karyotypes are given within the delineated and labelled boxes.



(a) LogRRatio, X and Pseudo-autosomal SNPs, Sample A

(b) BAlleleFreq X and Pseudo-autosomal SNPs, Sample A

Figure 5: An abnormal LogRRatio scan of probes from the X chromosome and the pseudoautosomal region for sample A, who is suspected to have an XY/XYY mosaic as detected by higher than expected Y chromosome intensity.

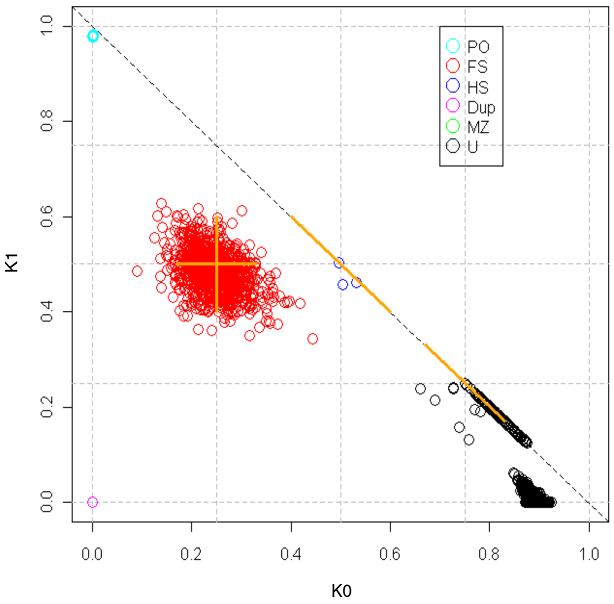


Figure 6: IBD coefficients to estimate relatedness. Each point represents a pair of participants and the diagonal line is $k0 + k1 = 1$, and the orange bars indicate the expected values of $k1 \pm 2$ standard deviations for full siblings (0.25 ± 0.08), half siblings (0.5 ± 0.10) and first cousins (0.25 ± 0.08). This plot shows all study pairs of participants with an estimated $KC > 1/32$, color-coded by inferred relationships. "PO" means parent-offspring, "FS" indicates full siblings, "HS" denotes half-sibling like relationships, "Dup" denotes duplicates, "MZ" indicates monozygotic twin pairs and "Unrel" means unrelated samples. The parent-offspring pairs are HapMap controls.

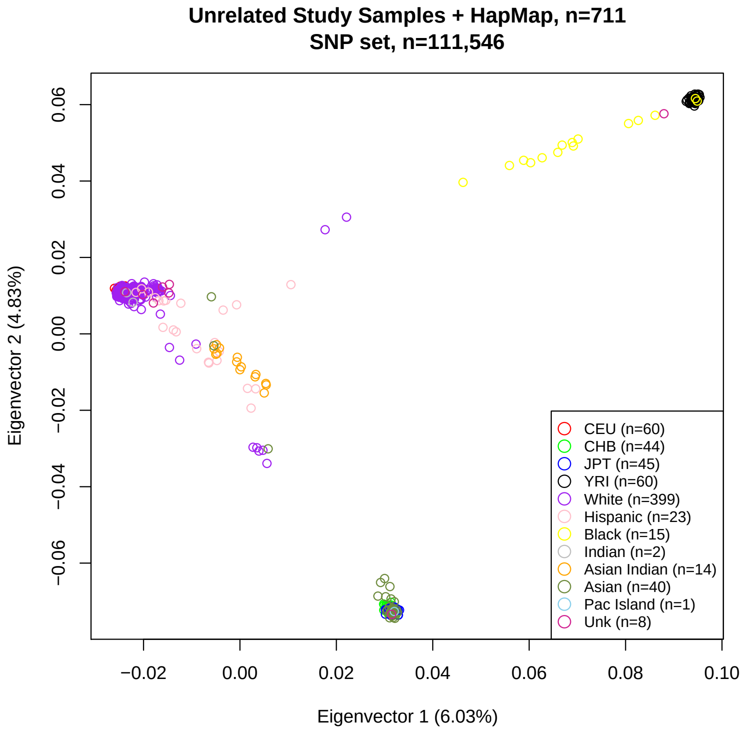


Figure 7: Principal components analysis of all unduplicated, unrelated study participants with 209 HapMap control founders using 111,546 SNPs. “CEU”, “CHB”, “JPT” and “YRI” indicate HapMap samples external to controls genotyped with the study participants. The study samples are color-coded by self-identified ethnicity. Numbers shown on axis labels are the percent variance accounted for by the respective eigenvector.

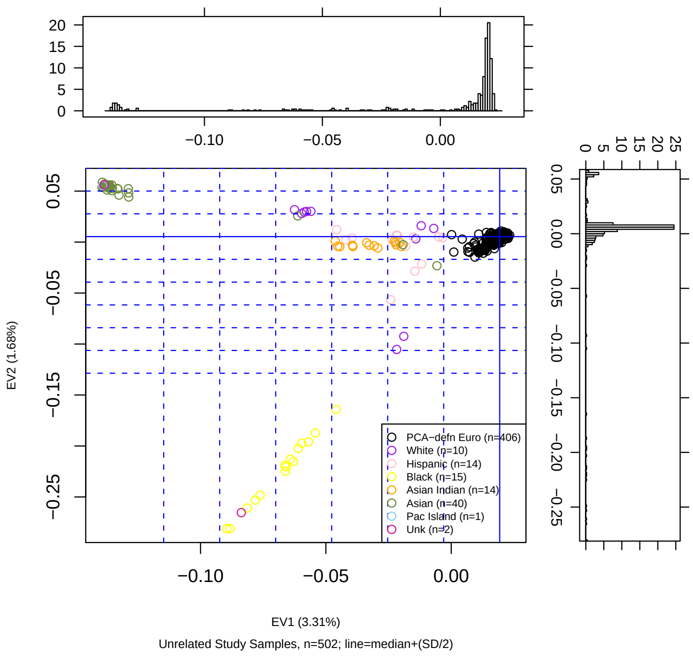
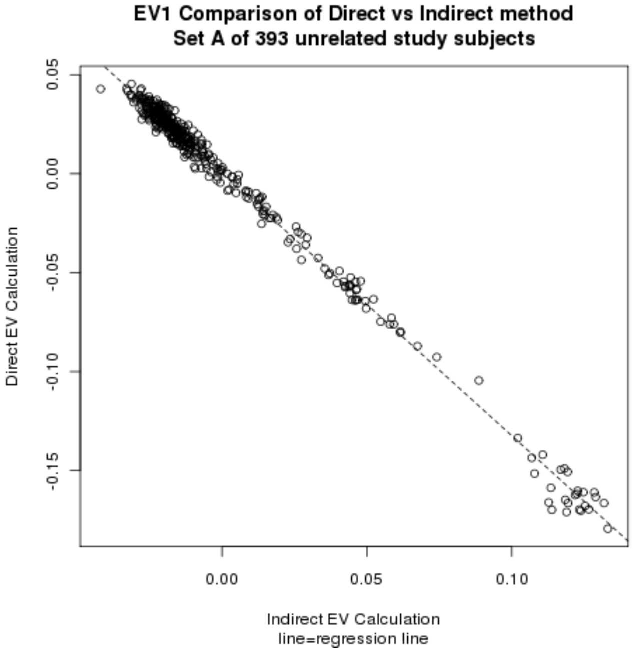
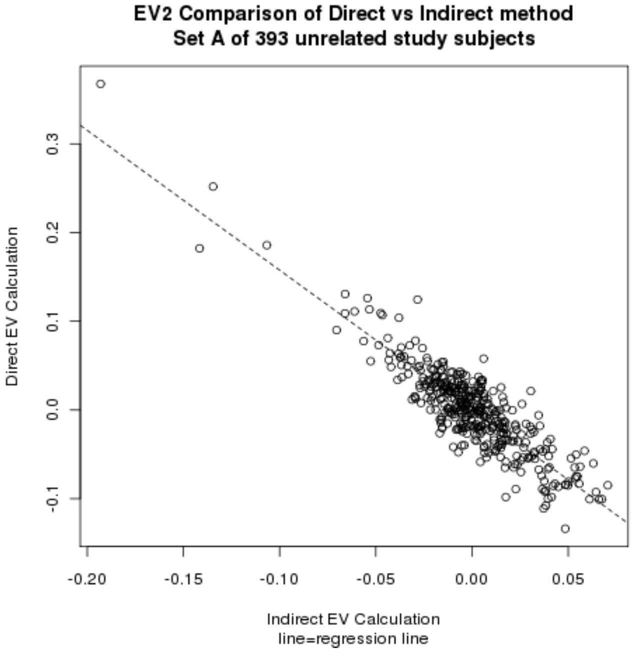


Figure 8: Principal components analysis of all unduplicated, unrelated study samples using 111,546 SNPs. The solid line is the median value for eigenvectors one and two, and the dotted lines mark each one-half SD. Samples are colored by self-identified ethnicity and those colored black are samples that lie within one-half SD from the median for eigenvectors one and two. These 406 black-colored samples comprise the homogeneous PCA-defined European sample set. Numbers shown on axis labels are the percent variance accounted for by the respective eigenvector.

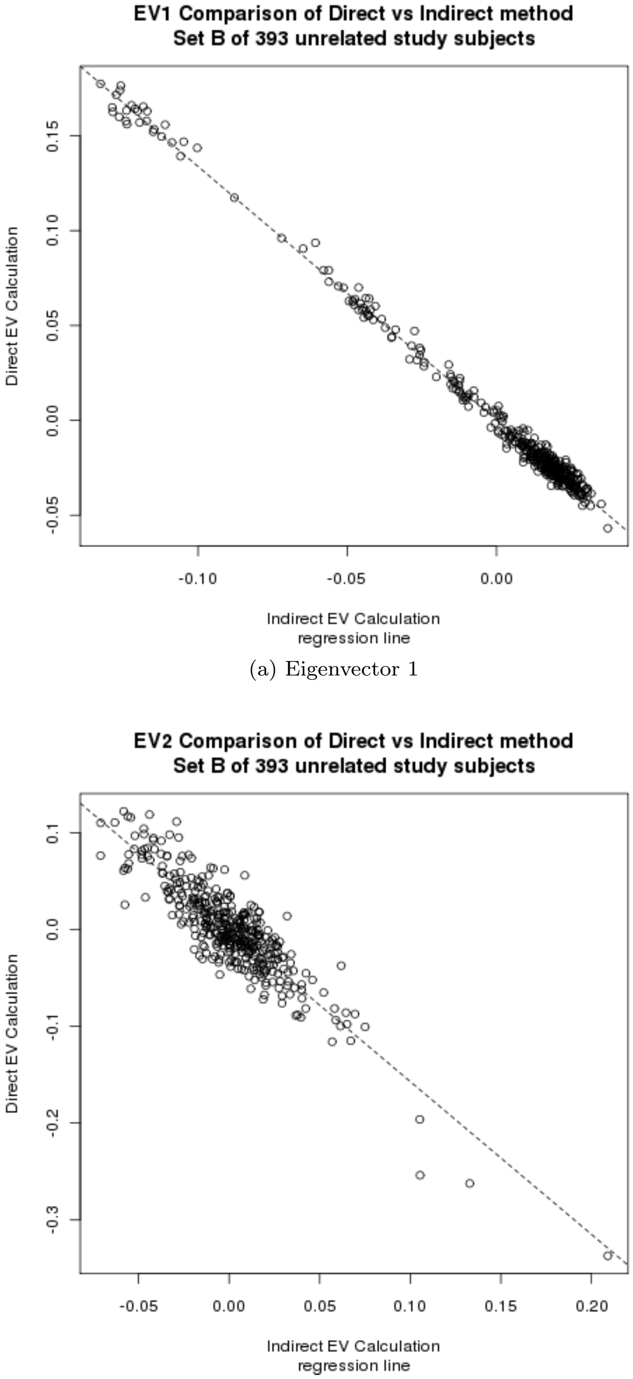


(a) Eigenvector 1



(b) Eigenvector 2

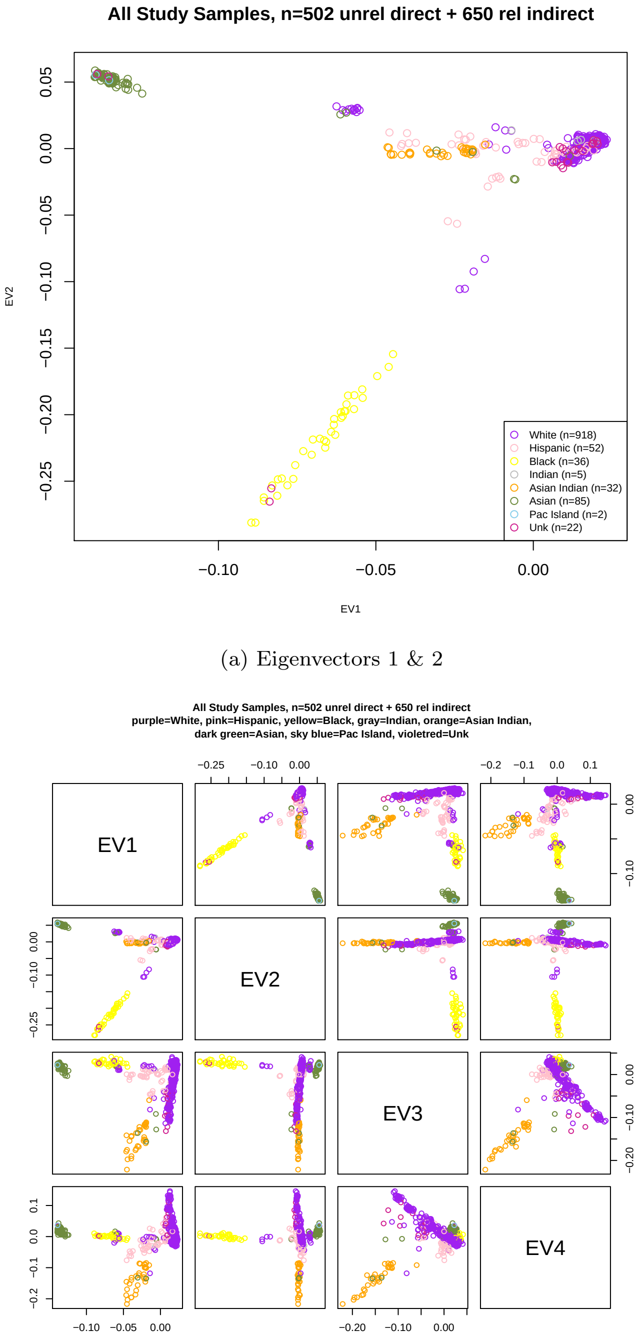
Figure 9: Comparison of eigenvector one for a randomly chosen set of 393 unrelated study subjects, indirect versus direct results.



(a) Eigenvector 1

(b) Eigenvector 2

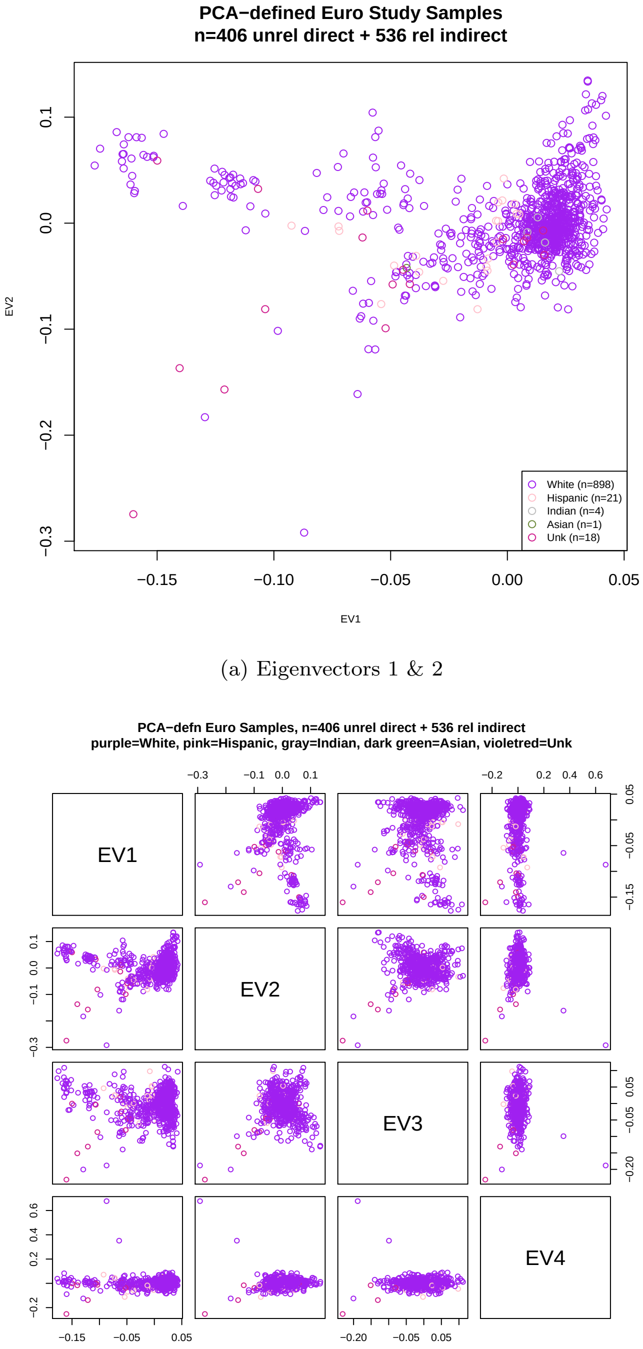
Figure 10: Comparison of eigenvector one for a randomly chosen set of 393 unrelated study subjects, indirect versus direct results.



(a) Eigenvectors 1 & 2

(b) Eigenvectors 1-4

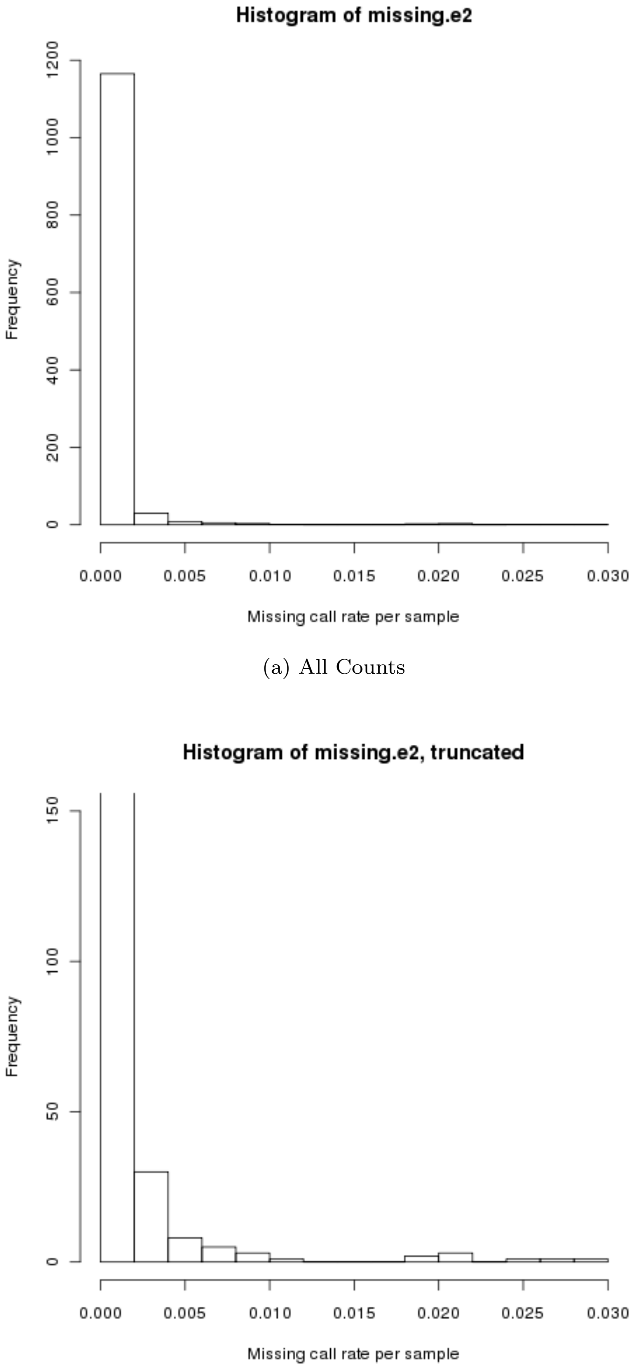
Figure 11: Principal components analysis calculated directly for 502 unduplicated, unrelated study samples using 111,546 SNPs. Eigenvectors were inferred for the remaining 650 study samples. Plotted are the results of the direct and indirect analyses.



(a) Eigenvectors 1 & 2

(b) Eigenvectors 1-4

Figure 12: Principal components analysis calculated directly for 406 unduplicated, unrelated PCA-defined European study samples using 103,464 SNPs. Eigenvectors were inferred for the remaining 536 family members. Plotted are the results of the direct and indirect analyses.



(a) All Counts

(b) Truncated

Figure 13: Histograms of the missing call rate per sample (*missing.e2*).

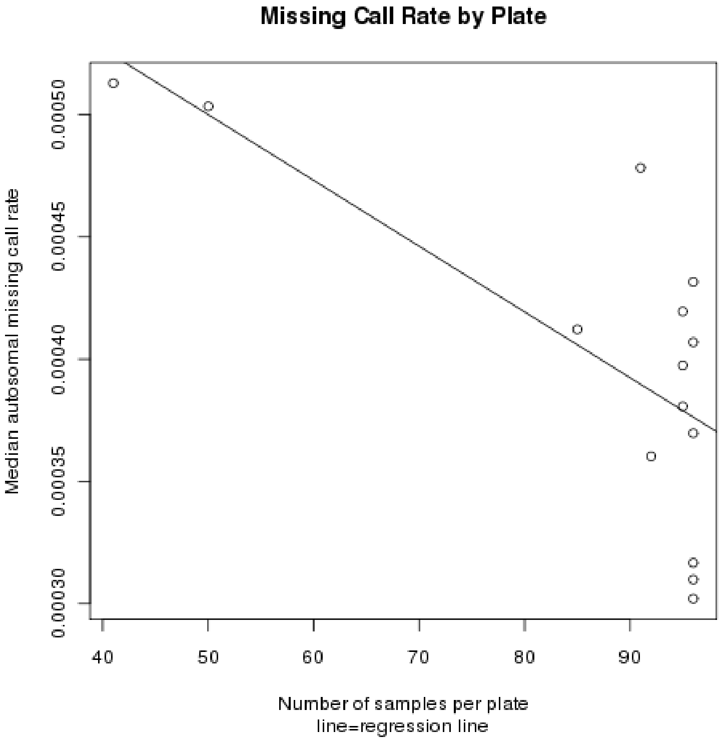


Figure 14: Median autosomal missing call rate versus number of samples per batch. Line plotted is the regression line.

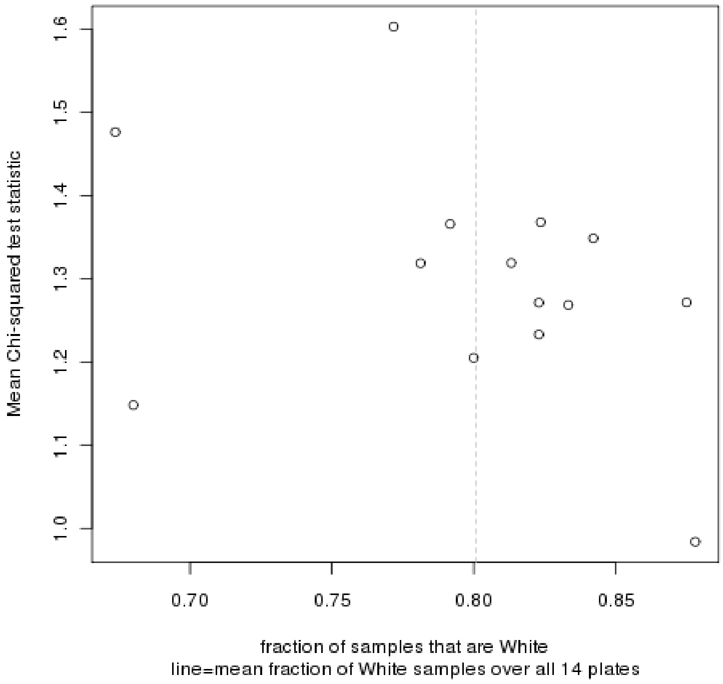


Figure 15: A test of allele frequency difference between each genotyping batch and a pool of the other batches, plotted as a function of the ethnic composition of the samples in the batch. The vertical dashed line is the mean composition.

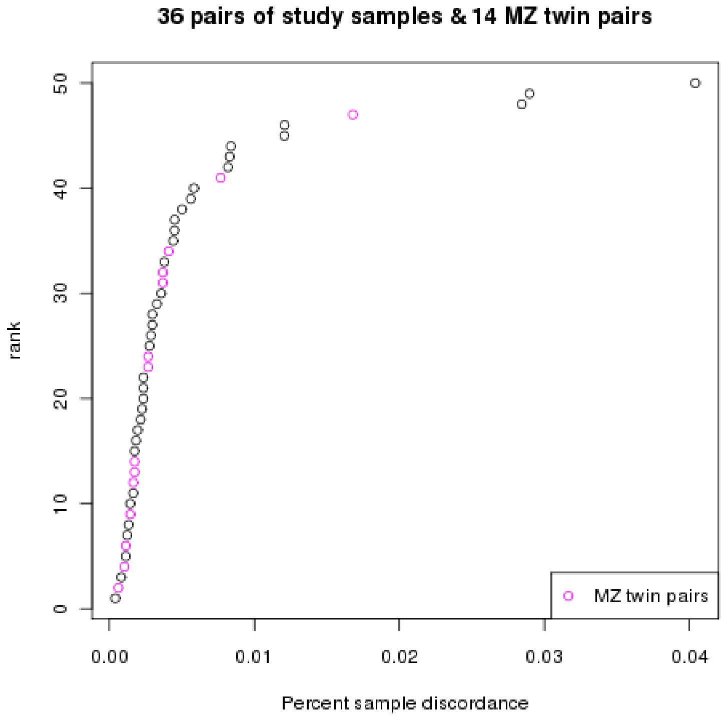
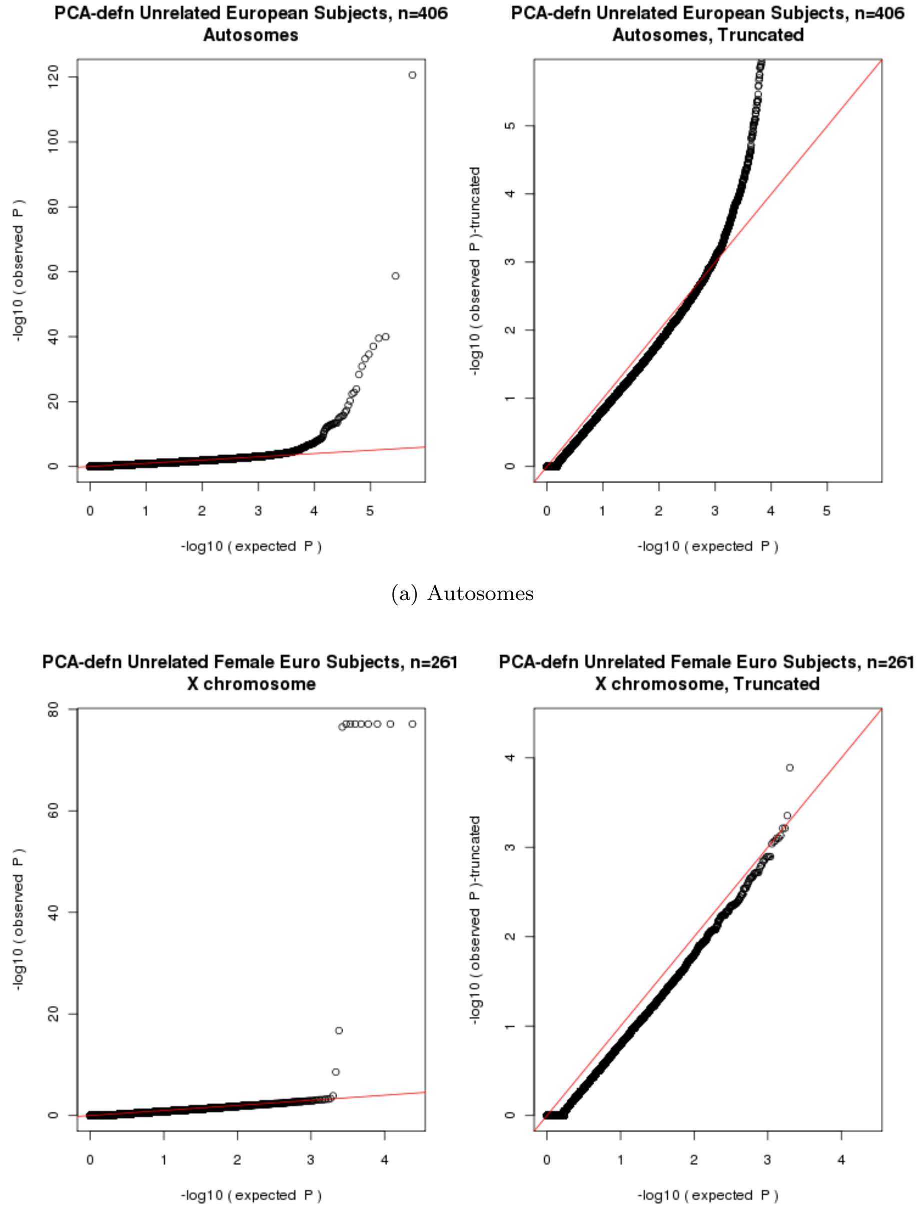


Figure 16: Increasing duplicate genotype discordance between duplicate samples of 36 study participants, along with 14 monozygotic twin pairs. The twin pairs are plotted in magenta. The median discordance is 0.0029%.



(a) Autosomes

(b) X Chromosome

Figure 17: Quantile-quantile plots of the p-values from the Hardy-Weinberg exact test run for unrelated study participants of primarily European ancestry.

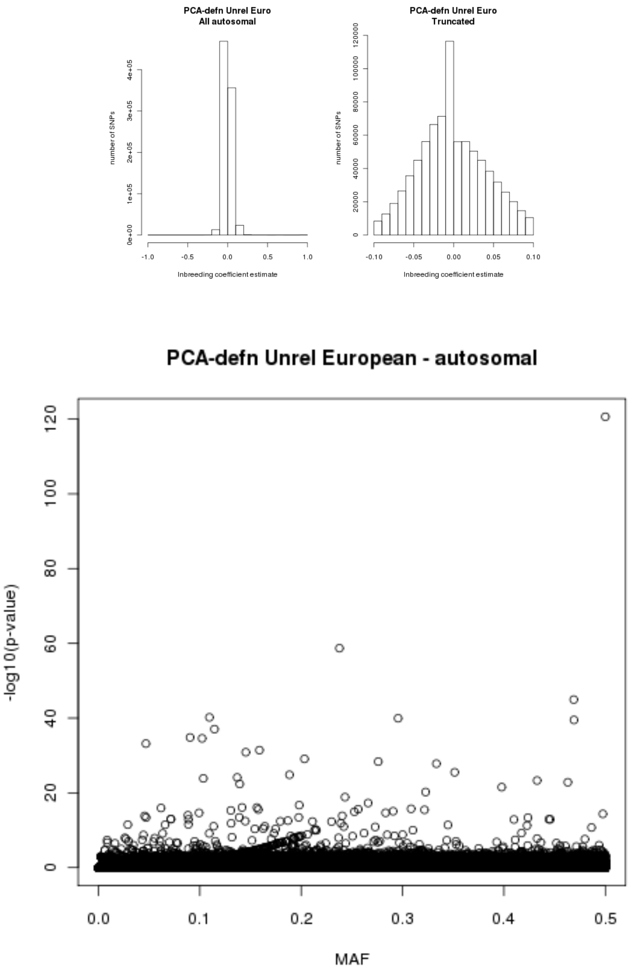


Figure 18: Hardy-Weinberg results. Figure 18a are histograms showing the distribution of the estimated inbreeding coefficient, where one is truncated to show the distribution roughly centering around zero. The range of the inbreeding coefficient is from -1 to 1 and the median value is $-2.469x10^{-3}$. Figure 18b shows the relationship of the p-value for the Hardy-Weinberg exact test to minor allele frequency.

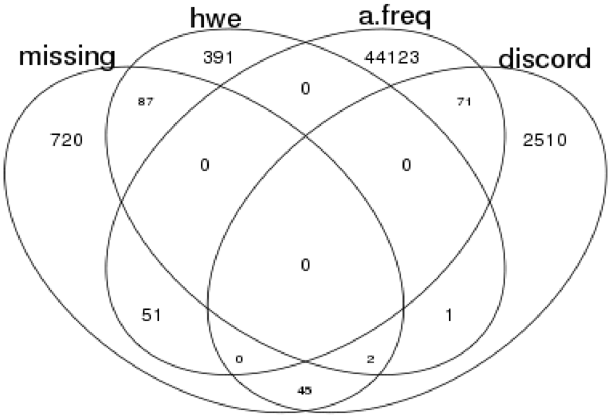


Figure 19: Venn diagram showing the overlap of various SNP filters. In the diagram,“missing” refers to SNPs filtered out with a missing call rate ≥ 2%, “hwe” shows the SNPs filtered out with a HWE p-value $< 10^{-4}$, “a.freq” refers to SNPs with allele frequency in the study samples $0 < a.freq < 1$, and “discord” refers to SNPs with > 0 discordant calls in 36 duplicate study sample pairs and 14 monozygotic twin pairs. No overlapping SNPs were detected by every filter.

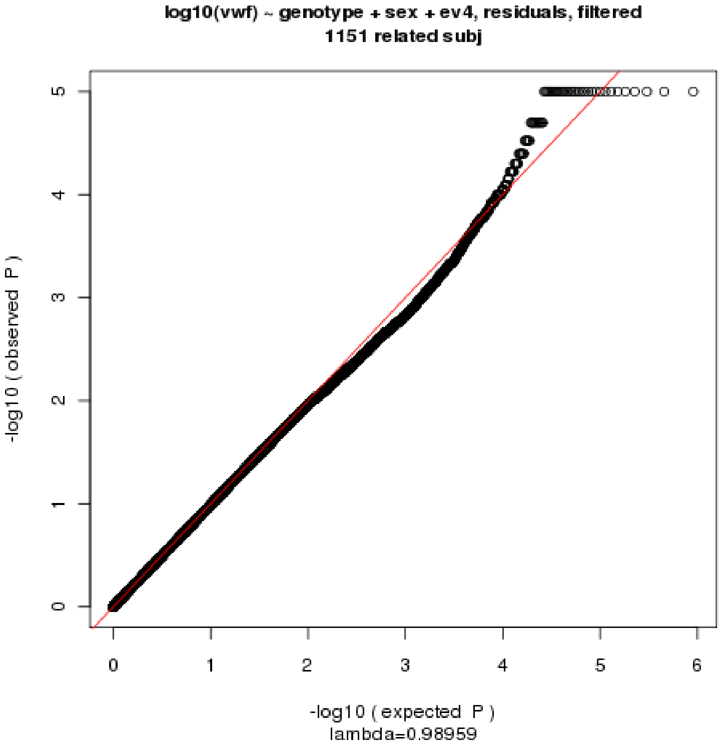


Figure 20: Quantile-quantile plots of p-values from the PLINK QFAM model (as described in Table 8). The corresponding genomic control factors are also described in the formerly mentioned table. The p-values are filtered as described in Table 1 for quality. The ceiling p-value is 10^{-5} because this test was run with 100,000 permutations.

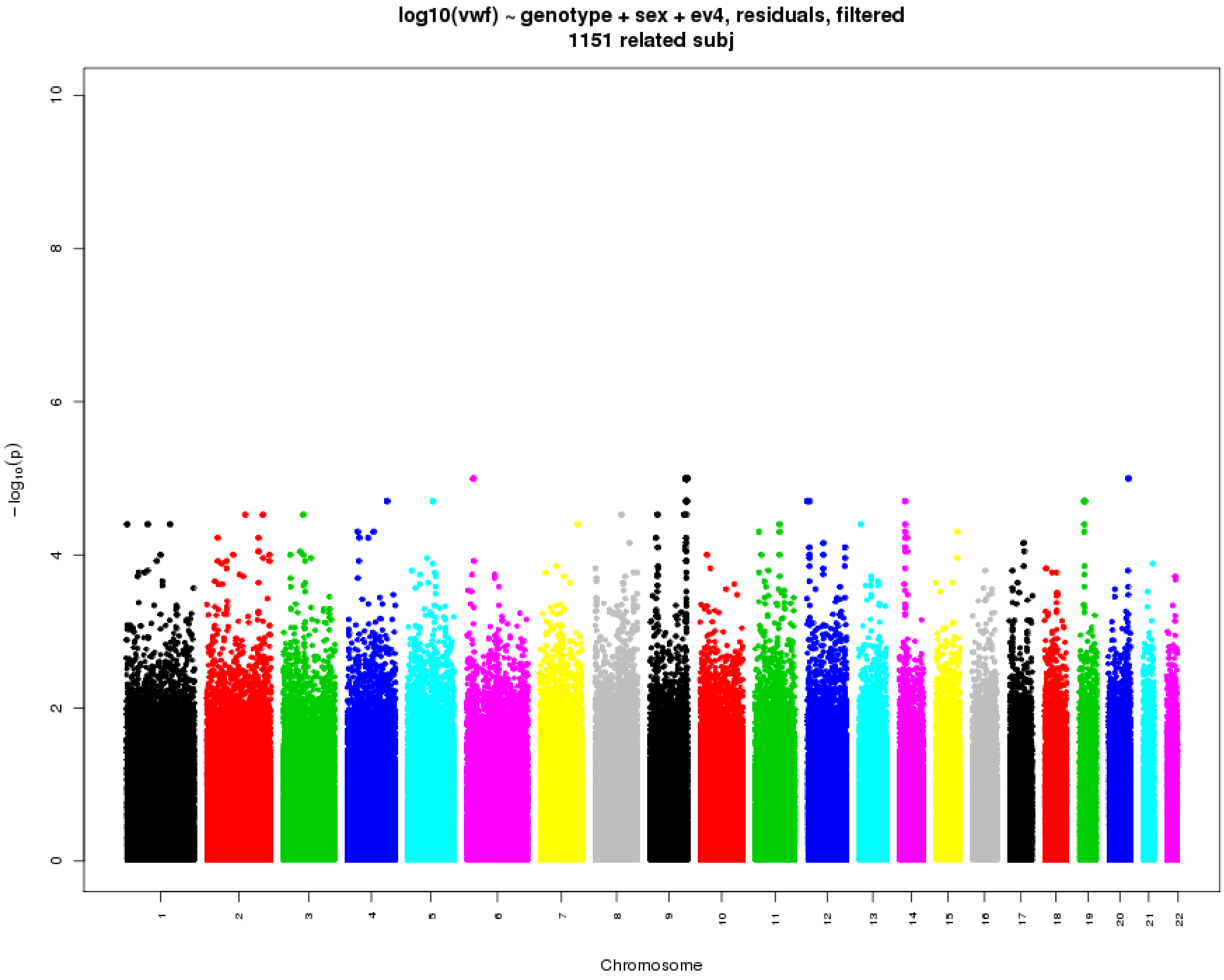


Figure 21: Manhattan plots for model A1 (Table 8). Plots are for p-values after applying the SNP quality filter (Table 1).

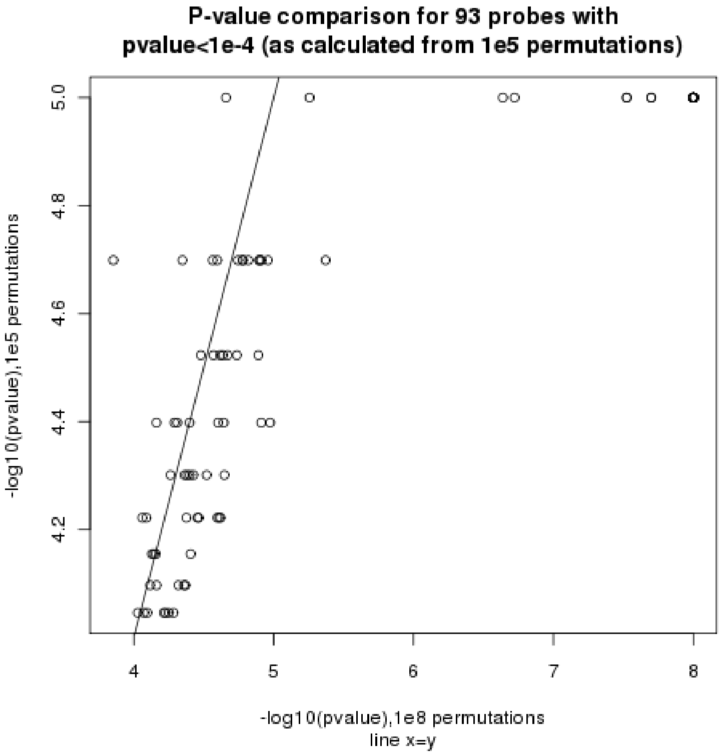


Figure 22: Top 93 p-values from PLINK QFAM association test, run using 1e5 permutations compared with p-values obtained from run using 1e8 permutations. Line plotted is x=y.

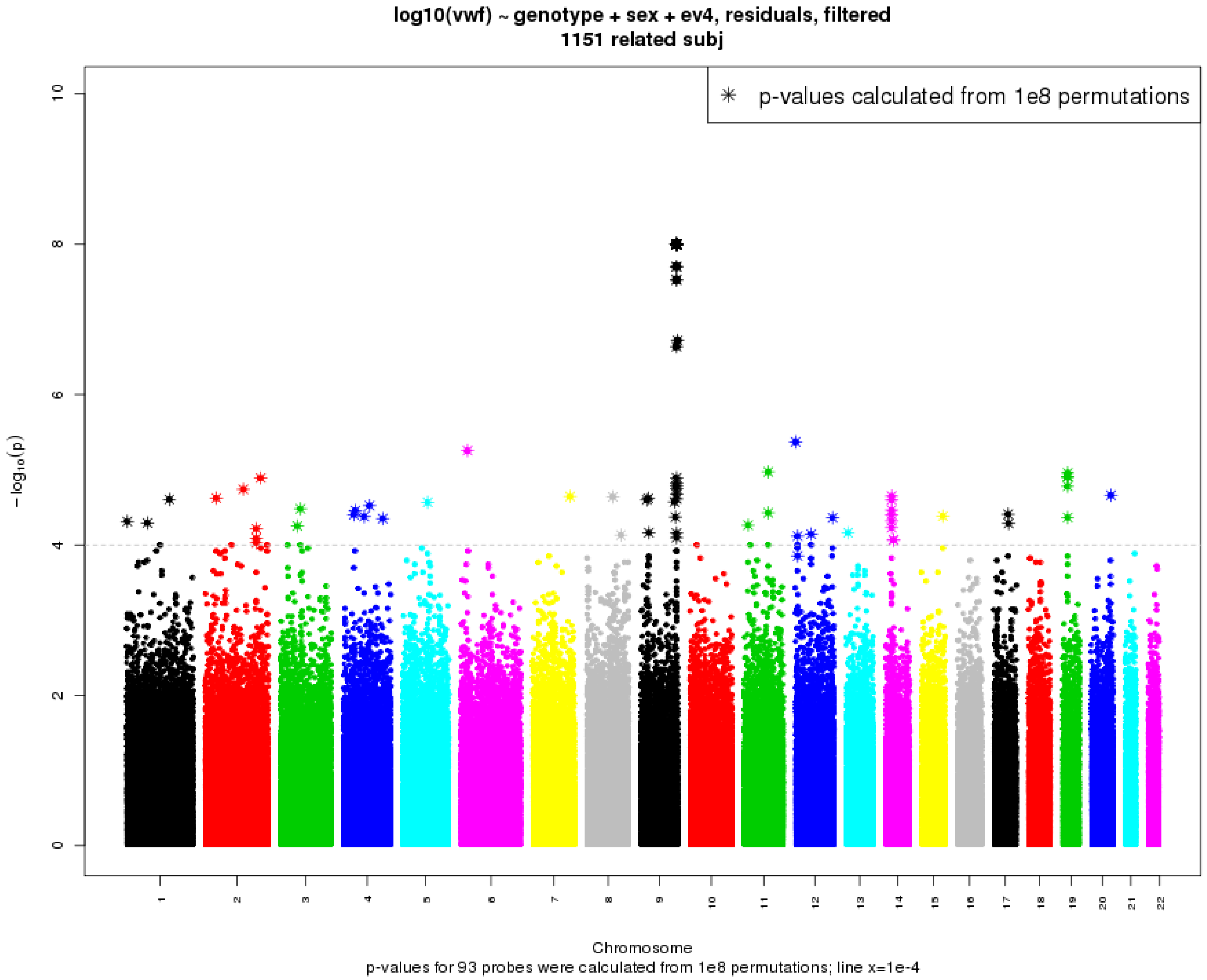


Figure 23: Manhattan plots for model A2 (Table 8). SNPs with the lowest p-values from run with 1e5 permutations are plotted with p-values calculated from 1e8 permutations. Plots are for p-values after applying the SNP quality filter (Table 1).

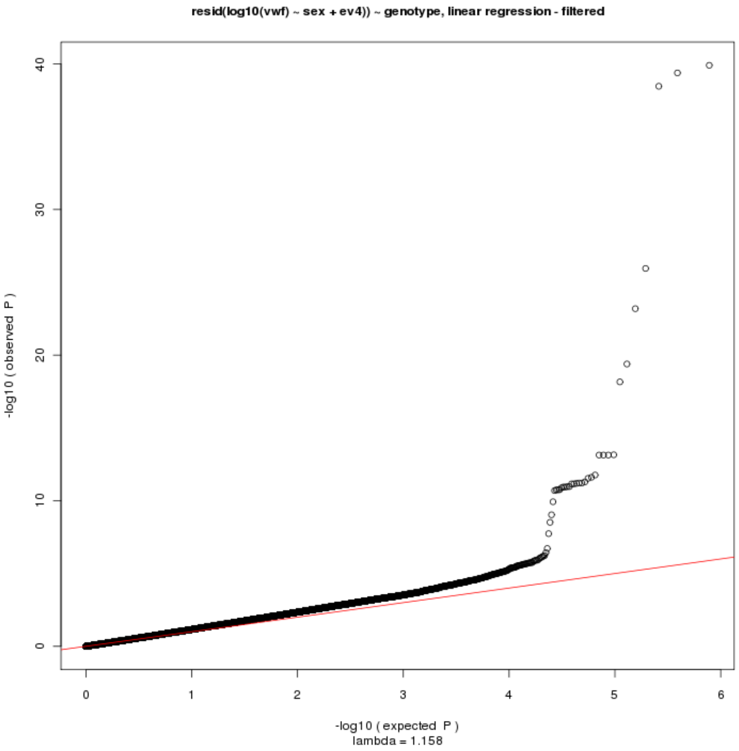


Figure 24: Quantile-quantile plots of p-values from linear association model (as described in Table 8). The corresponding genomic control factor is also described in the formerly mentioned table. The p-values are filtered as described in Table 1 for quality and MAF<0.02 filter. This test ignores relatedness.

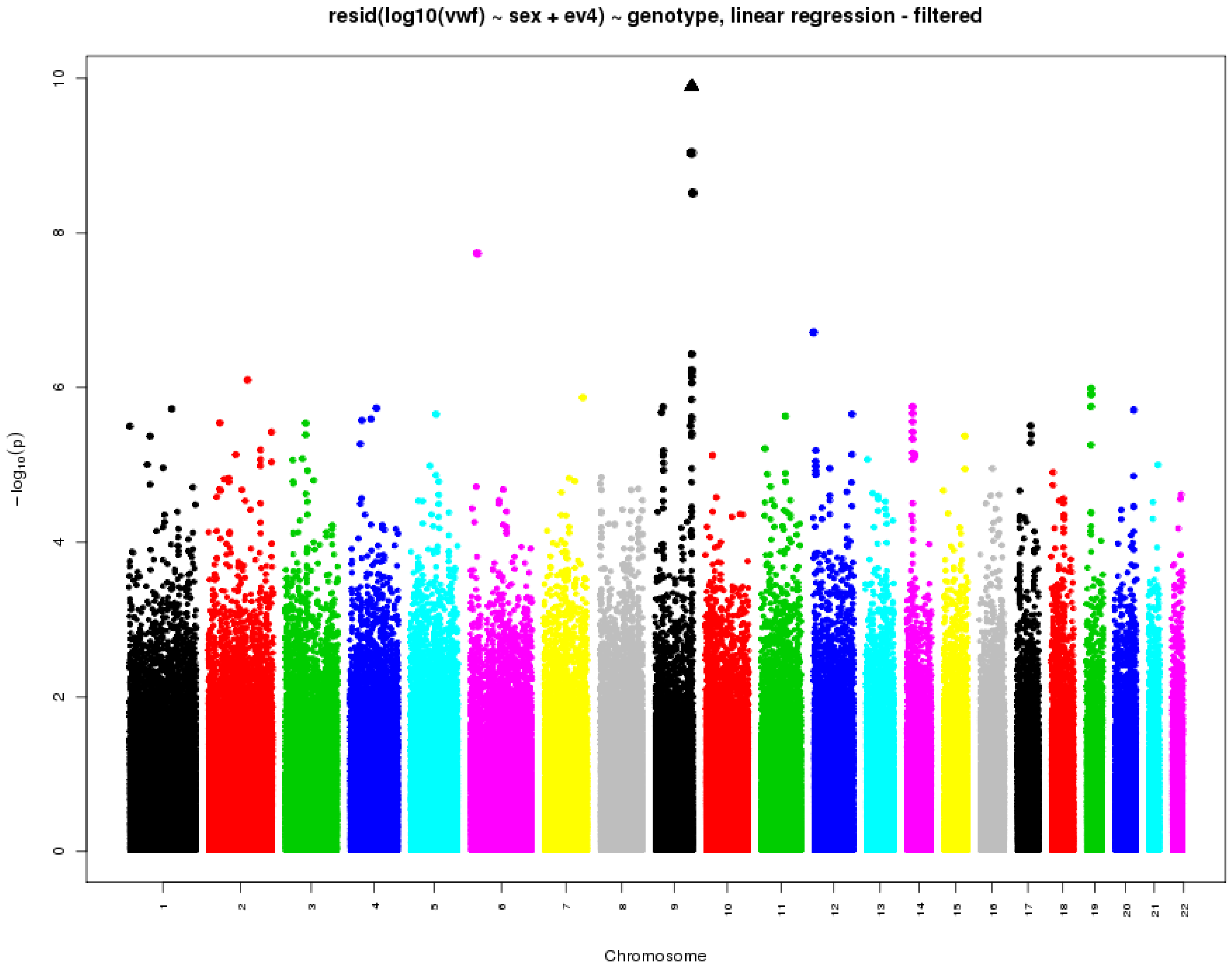


Figure 25: Manhattan plots for model A3 (Table 8). Plots are for p-values after applying the SNP quality filter (Table 1) and MAF < 0.02 filter. This test ignores relatedness.