Quality Control Report for Genotypic Data

University of Washington

March 5, 2012

Project: Health Retirement Study
Principal Investigator: David R. Weir, University of Michigan
Support: CIDR Contract # HHSN268200782096C and HHSN268201100011I
NIH Institute: NIA

Contents

List of Tables

List of Figures

1 Summary and recommendations for dbGaP users

A total of 12,507 study subjects were genotyped on the Illumina HumanOmni2.5-4v1 array. The median call rate is 99.7% and the error rate estimated from 336 pairs of study sample duplicates is 6×10^{-5}. Genotypic data are provided for all subjects and SNPs. We recommend selective filtering of genotypic data prior to analysis to remove large (> 10 Mb) chromosomal anomalies showing evidence of genotyping error and to remove whole samples with an overall missing call rate $> 2\%$. 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 Project overview

Since 1992, the Health and Retirement Study (HRS, a cooperative agreement between the National Institute on Aging (NIA) and the University of Michigan) has been the largest, most representative longitudinal study of Americans over age 50. Built on a national probability sample with oversamples of minorities, it is the model for a network of harmonized international longitudinal studies that monitors work, health, social, psychological, family and economic status, and assesses critical life transitions and trajectories related to retirement, economic security, health and function, social and behavioral function and support systems.

HRS study design and sample selection

The HRS is a nationally representative sample with a 2:1 oversample of African-American and Hispanic populations. The target population for the original HRS cohort includes all adults in the contiguous United States born during the years 1931–1941 who reside in households. HRS was subsequently augmented with additional cohorts in 1993 and 1998 to represent the entire population 51 and older in 1998 (b. 1947 and earlier). Since then, the steady-state design calls for refreshment every six years with a new six-year birth cohort of 51–56 year olds. This was done in 2004 with the Early Baby Boomers (b. 1948-53) and in 2010 with the Mid Boomers (b. 1954–59). The current sample includes over 26,000 persons in 17,000 households.

Core interview data are collected every two years using a mixed mode design, combining in-person and telephone interviews.

In 2006, a random one-half of the sample was pre-selected to complete an enhanced face-to-face (EFTF) interview, which included a set of physical performance tests, anthropometric measurements, blood and saliva samples, and a psychosocial self-administered questionnaire in addition to the core HRS interview. The sample was selected at the household-level. In 2008, an EFTF interview was conducted on the remaining half of the sample. Respondents who consented to the saliva collection in either 2006 or 2008 are included in this data release.

HRS phenotypic data

Phenotypic data are available on a variety of dimensions. Health measures include self-reported doctor-diagnosed disease and some aspects of treatment, including medications, health insurance and utilization, smoking, drinking, height, weight, physical function, family characteristics and interactions, income, wealth and financial management, and job conditions. It measures cognitive ability in several domains and depression. Innovative measures of risk tolerance and time preference have been used, as well as probabilistic expectations. The study is supplemented with administrative linkages to Medicare claims files providing diagnostic and utilization information, to the National Death Index, and to Social Security.

Beginning in 2006 the study added direct measures of physical function (grip strength, gait speed, balance, lung function), biomarkers of cardiovascular risk (blood pressure, total and HDL cholesterol, HbA1c, C-

reactive protein and cystatin-C, height, weight, and waist circumference), and greatly expanded measurement of psychological traits (e.g., big 5 personality measures, affect, sense of control) and social networks.

Performance on a cognitive test combining immediate and delayed word recall was selected as an example trait for the dbGaP data release. In the immediate word recall task the interviewer reads a list of 10 nouns to the respondent and asks the respondent to recall as many words as possible from the list in any order. After approximately five minutes of asking other survey questions, the respondent is asked to recall the nouns previously presented as part of the immediate recall task. The total recall score is the sum of the correct answers to these two tasks, with a range of 0 to 20.

Researchers who wish to link to other HRS measures not in dbGaP will be able to apply for access from HRS. A separate Data Use Agreement (DUA) will be required for linkage to the HRS data. See the HRS website (http://hrsonline.isr.umich.edu) for details.

3 Genotyping process

This study was genotyped in two phases. The first phase consists of DNA from buccal swabs collected in 2006 and extracted using the Qiagen Autopure method. The second phase consists of saliva samples collected in 2008 and extracted with Oragene. Although the two phases were genotyped separately, the raw data were clustered and called together. The samples were genotyped in batches corresponding to 96-well plates. Each plate contained from one to four HapMap controls, as well as an average of two study sample duplicates. Additionally, 30 Oragene duplicates were genotyped with the Buccal samples, and 30 Buccal duplicates were genotyped with the Oragene samples. The DNA samples were genotyped at the Center for Inherited Disease Research (CIDR) using the Illumina HumanOmni2.5-4v1 array and using the calling algorithm GenomeStudio version 2011.2, Genotyping Module 1.9.4 and GenTrain version 1.0. The SNP annotation used by CIDR is "HumanOmni2.5-4v1_D", but during data cleaning at the Genetics Coordinating Center of the University of Washington (UWGCC), the annotation was updated to "HumanOmni2.5-4v1_H." Versions D and H have the same SNP identifiers, probes and design sequences, but there are some differences in map position. Also, version D has two SNPs that are not found in version H. These two SNPs were retained with their version D annotations. (See file "SNP_annotation.csv" for more details.) Both versions use genome build 37/hg19.

4 Quality control process and participants

Genotypic data that passed initial quality control at CIDR were released to the Quality Assurance/Quality Control (QA/QC) analysis team at the UWGCC, the study investigator's team and dbGaP. These data were analyzed by all four groups and the results were discussed in periodic conference calls. Key participants in this process and their institutional affiliations are given in Appendix A. Analysis tools varied by group, but the results presented here were generated with the R packages *GWASTools*[1] and *SNPRelate*[2], unless indicated otherwise. The methods of QA/QC used here are described by Laurie et al. [1].

5 Sample and participant number and composition

In the following, the term "sample" refers to a DNA sample and, for brevity, "scan" refers to a genotyping instance (including genotyping chemistry, array scanning, genotype calls, etc.).

A total of 13,129 samples (including duplicates) from study subjects were put into genotyping production, of which 12,857 were successfully genotyped and passed CIDR's QC process (Table 2). The subsequent QA process identified 12 subjects with unresolved identity issues. Of these 12, seven were unexpected duplicates identified by CIDR prior to release, two were determined to have questionable identity by the SI based on

[1] http://http://www.bioconductor.org/packages/devel/bioc/html/GWASTools.html
[2] http://cran.r-project.org/web/packages/SNPRelate/index.html

their HRS IDs, one was a respondent dropped from the HRS sample, and one was found to be an unexpected relative of another subject (see Section 8). The set of scans to be posted include 12,845 study participants and 411 HapMap controls. The 12,845 study scans derive from 12,507 subjects and include 336 subjects with duplicate scans (334 subjects with 2 scans each and 2 subjects with 3 scans each). (Table 3). The 411 HapMap control scans derive from 88 subjects, of which 87 are replicated two or more times. The study subjects occur as 12,335 singletons and 84 families of two or three members each. The study families were discovered during the analysis of relatedness (Section 8). The HapMap controls include 25 trios (11 CEU, 4 MEX, and 10 YRI) as well as 13 singletons (1 CEU, 6 CHB, and 6 JPT).

6 Gender identity

To check gender identity, we look at both X chromosome heterozygosity and the means of the intensities of SNP probes on the X and Y chromosomes. The expectation is that male and female samples will fall into distinct clusters that differ markedly in X and Y intensities. Figure 1 shows that there are two main clusters, as expected, and no gender mis-annotations are apparent. There are tails of low Y intensity for males and low X intensity for females, which we have found is not unusual for elderly subjects. There are a number of sex chromosome anomalies, annotated in Figure 2, which are discussed further in Section 7.

7 Chromosomal anomalies

Large chromosomal anomalies, such as aneuploidy, copy number variations and mosaic uniparental disomy, can be detected using "Log R Ratio" (LRR) and "B Allele Frequency" (BAF) [2, 3]. LRR is a measure of relative signal intensity (log_2 of the ratio of observed to expected intensity, where the expectation is based on other samples). BAF is an estimate of the frequency of the B allele of a given SNP in the population of cells from which the DNA was extracted. In a normal cell, the B allele frequency at any locus is either 0 (AA), 0.5 (AB) or 1 (BB) and the expected LRR is 0. Both copy number changes and copy-neutral changes from biparental to uniparental disomy (UPD) result in changes in BAF, while copy number changes also affect LRR.

To identify aneuploid or mosaic samples systematically, we used two methods. For anomalies that split the intermediate BAF band into two components, we used Circular Binary Segmentation (CBS) [4] on BAF values for SNPs not called as homozygotes. For heterozygous deletions (with loss of the intermediate BAF band), we identified runs of homozygosity accompanied by a decrease in LRR. All sample-chromosome combinations with anomalies greater than 10 Mb were verified by manual review of BAF and LRR plots.

Figure 3 shows BAF/LRR plots for chromosome 1 in sample A. This chromosome shows a normal pattern, with LRR centered at 0 and all three BAF bands at 0, 0.5, and 1 (corresponding to AA, AB, and BB genotypes). Figure 4 shows BAF/LRR plots for chromosome 16 in the same sample, which shows an abnormal pattern, i.e., a split in the heterozygous band wide enough to cause genotyping errors, with some heterozygotes evidently called as homozygotes. This anomaly appears to be a copy-neutral region of uniparental disomy, most likely caused by mitotic recombination.

Figure 5 shows BAF/LRR plots for chromosome X in Female B. This chromosome shows a pattern consistent with trisomy, i.e., a split in the heterozygous band with positions at about 1/3 and 2/3, and elevated LRR. Therefore, we consider the blood cell karyotype of this subject to be 47, XXX.

Figure 6 shows BAF/LRR plots for chromosome X in Female C. In this chromosome, it appears that the split in the heterozygous band is so wide that it has merged with the homozygous bands. This is consistent with an XO or XX/XO mosaic karyotype.

Figure 7 shows BAF/LRR plots for chromosome X in Female D. This chromosome has two separate anomalies with different degrees of severity. The anomaly on the left is so wide that it appears to result in genotyping errors, with some heterozygotes called as homozygotes. The anomaly on the right is less severe, with an increase in missing calls but not genotyping errors.

All large chromosome anomalies (> 10 Mb), regardless of whether they are recommended for filtering, are provided in the file "chromosome_anomalies.csv." The indication of which SNPs should be filtered was made based on manual review of the BAF/LRR plots to determine which anomalies are associated with genotyping errors. Of 136 large autosomal anomalies, we recommend filtering 29.

Of the sex chromosome anomalies shown in Figure 2, the following decisions were made regarding filtering:

- Of the seven XXX anomalies, only one was severe enough to filter (shown in Figure 5).
- Of the 14 XX/XO samples, 8 are recommended for filtering.
- For the 5 XXY samples (males with two copies of the X chromosome), filter the X.
- Filter the Y chromosome for 43 males with Y intensity < 0.5 (approximately the top of the Y intensity distribution for females).

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

$$KC = \frac{1}{2}k2 + \frac{1}{4}k1 \tag{1}$$

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 [5, 6, 7, 8] and have been extended by the UWGCC to the three IBD coefficients [9].

Given the large number of samples and ethnic diversity in this study, three ethnic groups were defined using results from an initial round of Principal Component Analysis (PCA). Section 9 describes the procedure for PCA in more detail. Figure 13 shows the distribution of the races and ethnic groups with eigenvectors 1 and 2. (Although this plot shows the results of PCA with only unrelated study subjects, the plot with all subjects was essentially identical.) The study samples were divided into three categories for the purposes of IBD: "African-American" with EV1 >= 0.08, "European" with EV1 < 0.008 and EV2 < 0.006, and "Hispanic/Asian" with EV2 >= 0.006.

IBD coefficients were estimated using autosomal SNPs and the Method of Moments procedure used by PLINK [10], but implemented in R using the package *SNPRelate*. The SNPs were selected at random from all SNPs that are autosomal, non-monomorphic in the relevant ethnic group, and had missing call rate < 5%, with the constraint that no two SNPs are closer than 15 kb. The exact SNPs selected varied with ethnic group, but ~125,000 SNPs were used in each case.

Figures 8, 9, and 10 show all relationship pairs with $KC > 1/32$, which is half the expected value for first cousins, for the European, African-American, and Hispanic/Asian ethnic groups, respectively. The 99.6% prediction ellipse is shown for the full siblings and the 99.6% prediction intervals are shown for half-sibling/avuncular/grandparent-granchild and first cousin relationships (though the HS-like and FC intervals overlap at this prediction level). All expected study sample duplicates were observed. No genetic relatedness was annotated *a priori*, but we observed 90 unexpected relationships. These were categorized into relationship types using 99.6% prediction ellipse/intervals, as shown in the figures. There were no unexpected duplicates. There were 45 full sibling pairs, 20 half-sib-like pairs (half-sib/avuncular/grandparent-grandchild) and 25 parent-offspring pairs. We did not attempt to assign first-cousin relationships because we believe that the IBD estimates are not sufficiently reliable to distinguish between first cousins and unrelated subjects.

In one of the parent-offspring pairs, the daughter had been incorrectly annotated as being a spouse of the father. Since the HRS group was confident of the father's identity but not the daughter's, the daughter's sample will not be posted on dbGaP.

Because of the ambiguity in half-sib-like relationships, we were unable to specify a pedigree structure for the study subjects. Nevertheless, we defined families so that each family includes all pairs of subjects connected by a $KC > 0.1$ (the half-sib-like expectation minus 2 SD), and $k1 > 0$. This procedure resulted in 80 families of two members each and 4 families of three members each. The IBD coefficient estimates for these families are provided in the file "Kinship_coefficient_table.csv." *For an analysis that assumes all participants are unrelated, we recommend selecting one subject from each family unit. See Appendix B.*

9 Population structure

To investigate population structure, we use principal components analysis (PCA), essentially as described by Patterson et al. [11], but implemented in R (*SNPRelate* package). We use PCA for two purposes: to identify population group outliers and to provide sample eigenvectors as covariates in the statistical model used for association testing to adjust for possible population stratification.

We and others [12] have shown that it is often necessary to perform linkage disequilibrium (LD)-based or other pruning of the SNPs to be used for PCA, in order to avoid having sample eigenvectors that are determined by small clusters of SNPs at specific locations, such as the LCT, HLA or 8p23.1 (which contains a polymorphic inversion) [12]. Therefore, the SNPs used for PCA were selected by LD pruning from an initial pool consisting of all autosomal SNPs with a missing call rate $< 5\%$ and minor allele frequency (MAF) $> 5\%$. In addition, the 2q21 (LCT), HLA, 8p23, and 17q21.31 regions were excluded from the initial pool. The LD pruning process, using all unrelated study subjects, selected 154,644 SNPs with all pairs having $r^2 < 0.1$ in a sliding 10 Mb window.

Initially, we performed PCA on all study subjects and a set of 1230 HapMap controls, which were genotyped on the Illumina Human1M. The set of SNPs for this PCA was more limited, starting with only 492,164 SNPs in common between this array and the Omni2.5, and excluding any SNPs with a discordance between HapMap controls genotyped along with the study samples and those in the external HapMap data set. After LD pruning, 96,134 SNPs were selected.

These HapMap controls include subjects of many ancestries (ASW, CEU, CHB, CHD, GIH, JPT, LWK, MEX, MKK, TSI, and YRI). Figure 11 shows a plot of the first two eigenvectors from this analysis. Given the large number of subjects, Figure 12 shows plots with subsets of points for clarity. The self-identified race of the study subjects correlates well with the HapMap populations (Figure 12a). The self-identified Mexican-Americans match the position of the MEX HapMaps (Figure 12b). The African-Americans stretch from the HapMap Africans toward the HapMap Europeans, while the Whites cluster around the Europeans (Figure 12c). For several of the subjects in which self-identified race (black or white) strongly contrasts with the results of PCA, the self-identified race was corrected after review by the HRS group. The Asians fall into two categories, one corresponding with the CHB/CHD/JPT (Chinese/Japanese) group and one with the GIH (South Asian) group (Figure 12d).

We performed another PCA analysis using unrelated study subjects (based on family definitions from Section 8) without HapMap controls. Figure 13 shows the first two eigenvectors from this analysis, which looks very much like the results of the PCA including HapMaps. Pairwise plots of the first four eigenvectors are shown in Figure 14. This analysis was used to provide covariates for the preliminary association tests and the results are given in the file "Principal_Components.csv."

To determine whether the LD-pruning effectively prevented the occurrence of small clusters of SNPs that are highly correlated with a specific eigenvector, we examine plots of the correlation of each SNP with each eigenvector. These plots are similar to GWAS "Manhattan" plots except that the Y-axis has the SNP-eigenvector correlation rather than an association test p-value. Figure 15 shows these plots for the first 8 eigenvectors. Eigenvector 4 shows small clusters of SNPs on chromosomes 2 and 6, corresponding to the LCT and HLA regions, despite the fact that SNPs in these regions were not used in the PCA calculation.

To determine which eigenvectors might be useful covariates to adjust for population stratification in

association tests, we examine the scree plot for the PCA (Figure 16) and the association of each eigenvector with total recall score. (Table 8). The scree plot shows that the fraction of variance accounted for is small and approximately level after the first two components. However, the association tests indicate a significant relationship between total recall score and the sixth eigenvector ($p = 0.00155$). Therefore, we selected eigenvectors 1–6 for use as covariates in the preliminary association tests described in Section 20.

10 Missing call rates

Two missing call rates were calculated for each sample and for each SNP in the following way (and provided in files "SNP_analysis.csv" and "Sample_analysis.csv" on dbGaP). (1) *missing.n1* is the missing call rate per SNP over all samples (including HapMap controls). (2) *missing.e1* is the missing call rate per sample for all SNPs with *missing.n1*$< 100\%$. (3) *missing.n2* is the missing call rate per SNP over all samples with *missing.e1*$< 5\%$. In this project, all samples have *missing.e1*$< 5\%$, so *missing.n1*=*missing.n2*. (4) *missing.e2* is the missing call rate per sample over all SNPs with *missing.n2*$< 5\%$.

In this study, the two missing rates by sample are very similar, with median values of 0.00266 (*missing.e1*) and 0.00218 (*missing.e2*). Figure 17 shows the distribution of *missing.e1*. All samples have a missing rate less than 4%.

The two missing call rates by SNP are identical. Table 5 gives a summary of SNP genotyping failures and missingness by chromosome type. For SNPs that passed the genotyping center QC, the median value of *missing.n1* is 0.00113 and 91.9% of SNPs have a missing call rate $< 1\%$.

We recommend filtering out samples with a missing call rate $> 2\%$ *and SNPs with a missing call rate* $> 2\%$.

Since this study was genotyped in two phases, we tested for missing call rate differences between the two rounds of genotyping and by DNA source (which is not exactly correlated due to the cross-type duplicates that were run in each round). Figure 18 shows the results of linear regression of log_{10}(*missing.e1*) on both genotyping round and DNA source, with the cross-type duplicates from each round shown individually. While there are no significant missing call rate differences in DNA source (Figure 18a) or genotyping round (Figure 18b) when all samples are considered, there are differences when considering only the duplicate samples of each type run in both rounds (Figures 18c and 18d). *Although the total recall score is not related to DNA source or genotyping round, we suggest that investigators using other phenotypes make sure that those phenotypes are not associated with DNA source.* A missing call rate difference between cases and controls can lead to spurious associations, since missingness is often nonrandom [13].

11 Batch effects

The samples were processed together in batches consisting of complete or partial 96-well plates. There is a highly significant variation among batches in log_{10} of the autosomal missing call rate ($p < 2 \times 10^{-16}$), but all plates have a low mean missing call rate (Figure 19).

Another way to detect genotyping plate effects is to assess the difference in allelic frequencies between each plate and a pool of the other plates. We calculated the odds ratio from Fisher's exact test for each SNP and each plate and then averaged these statistics over SNPs, using only study samples. The mean odds ratio was calculated as $1/min(OR, 1/OR)$. This statistic is a measure of how different each plate is from the other plates. Figure 20 shows the mean odds ratio for all plates except two outliers with very small sample size ($n = 2$ and 3). The remaining outliers with $mean(OR) > 1.8$ have 16 and 22 samples per plate. We concluded that there are no problematic plate effects.

12 Sample quality

We examined BAF/LRR plots for evidence of sample contamination (more than 3 BAF bands on all chromosomes) and other artifacts. For this we examined scans that are high or low outliers for heterozygosity,

high outliers for BAF standard deviation (or non-homozygous genotypes), and high outliers for relatedness connectivity (the number of samples to which a sample appears to be related with kinship coefficient $< 1/32$).

We encountered two types of sample quality issues during the QA/QC process. The first is a repeatable pattern of extreme LRR waves across all chromosomes. (Figure 21). This pattern was observed in many samples with a varying degree of severity, and appears to be related to concentration. All of the most extreme cases were Oragene samples, and nearly all of them had low concentration. The genotypes appear to be called correctly with a low missing call rate overall, so these samples were kept in the data set. Approximately 45 samples had a very marked effect, and many others had more mild effects of a similar type. However, a high level of noise in the sample data has a tendency to mask the LRR wave effect, so in the case of noisy samples it is hard to determine the severity of the waves.

The second type of sample quality issue only affected the X chromosome, and has been seen in Buccal samples in other studies. (Figure 22). Note that the two bands of missing calls for all SNPs are not specific to this type of artifact, rather, they are a result of pseudoautosomal SNPs being incorrectly annotated on the X chromosome. There were three samples with this quality issue, and in this case we recommend that the X chromosome be filtered. These samples are included in the "chromosome_anomalies.csv" file.

13 Duplicate sample discordance

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 two error rates, 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 median discordance rate over all sample pairs is 1.1×10^{-4}, a rough estimate of the mean error rate is 6×10^{-5} errors per SNP per sample, indicating a high level of reproducibility.

Duplicate discordance estimates for individual SNPs can be used as a SNP quality filter. The challenge here is to find a level of discordance that would eliminate a large fraction of SNPs with high error rates, while retaining a large fraction with low error rates. The probability of observing $> x$ discordant genotypes in a total of n pairs of duplicates can be calculated using the binomial distribution. Table 6 shows these probabilities for $x = 0$–7 and $n = 428$. Here we chose $n = 423$ to correspond to the number of pairs of duplicate samples for both study and HapMap control samples. *We recommend a filter threshold of* > 4 *discordant calls because this retains* $> 99\%$ *of SNPs with an error rate* $< 10^{-3}$*, while removing* $> 92\%$ *of SNPs with an error rate* $> 10^{-2}$. This threshold eliminates 2896 SNPs.

Figure 23 summarizes the concordance by SNP, binned by MAF. Figure 23a shows the number of SNPs in each MAF bin. Figure 23b shows the correlation of allelic dosage (r^2), which is greater for SNPs with higher MAF. Figure 23c shows the overall concordance, which is higher for SNPs with low MAF because these SNPs are most likely to be called as homozygous for the major allele and thus be concordant by chance. Figure 23d shows the minor allele concordance, which is the concordance among sample pairs with at least one copy of the minor allele (i.e., matches of major homozygotes excluded). This concordance measure is more reflective of true genotyping concordance for low-MAF SNPs and the distribution is very similar to the correlation. The minor allele concordance falls off sharply for MAF< 0.01.

14 Mendelian errors

Mendelian errors were analyzed in the 25 trios of HapMap control subjects. Only 0.9% of SNPs have any Mendelian errors and 2237 SNPs have more than one error. *We recommend filtering out SNPs with more than one Mendelian error to avoid removing SNPs with an error in just one trio, which might be due to copy number variation or other chromosomal anomaly.*

15 Hardy-Weinberg equilibrium

We calculated an exact test of Hardy-Weinberg equilibrium (HWE) using study subjects who are (1) unrelated, (2) have missing call rate $< 2\%$, (3) self-identified white and (4) fall within 1 SD of all self-identified whites for eigenvectors 1 and 2 in the PCA of all unrelated study subjects. Figure 24 shows quantile-quantile (QQ) plots for this HWE test.

A second HWE test used study subjects who are (1) unrelated, (2) have missing call rate $< 2\%$, (3) self-identified black and (4) fall within 2 SD of all self-identified whites for eigenvector 1 and 1 SD for eigenvector 2 in the PCA of all unrelated study subjects. Figure 25 shows quantile-quantile (QQ) plots for this HWE test.

In both populations, autosomal SNPs deviate from expectation just below 0.01, while the X chromosome SNPs deviate from expectation between 0.01 and 0.001. The X versus autosomal difference has been observed in many other studies. The reason(s) for it are not clear, but appear to be unrelated to sample size, since the difference generally is observed even when only females are analyzed for autosomes.

Deviations from HWE due to population structure is expected to result in an excess of homozygotes or a positive inbreeding coefficient estimate, calculated as 1−(number of observed heterozygotes)/(number of expected heterozygotes). Figure 26 shows the distribution of the inbreeding coefficient estimates for all autosomal SNPs. The distributions are roughly symmetrical with mean = 0.0023 and median = −0.00017 (European), mean = 0.0019 and median = −0.0018 (African). There does not appear to be an excess of positive coefficients. We conclude that most deviations from HWE result from genotyping artifacts, rather than population structure.

Although the QQ plots show deviation of observed from expected p-values for autosomal SNPs between 0.001 and 0.01, *we suggest using a filter threshold of* $p = 0.0001$ *because examination of cluster plots reveals good plots for many assays with p-values* > 0.0001. This threshold is rather subjective, but we are reluctant to recommend a higher threshold that would eliminate many good SNP assays.

16 Minor allele frequency

Figure 27 shows the distribution of minor allele frequency (MAF) for all unrelated study subjects. The percentage of all SNPs with MAF $< 1\%$ is 24.5% for the autosomes and 17.2% for the X chromosome.

17 Duplicate SNP probes

The Illumina HumanOmni2.5-4v1 array, with version H annotation, has 9,769 pairs of SNPs that occur in duplicate, as indicated by duplicated AlleleA_ProbeSeq, TopGenomicSeq and/or genomic position. Generally one member of the pair is from dbSNP, while the other is from 1000 Genomes. Most of these pairs have a very high level of concordance across the 12,845 study samples. The numbers of pairs of duplicate SNPs with various levels of discordance are given in Table 7, along with the probability of observing each level given an assumed error rate (estimated as 6×10^{-5} over all SNPs for this study). A high level of discordance may indicate that the SNP has a high error rate or that the two members of the pair may not be assaying the same SNP.

We recommend filtering out both members of each SNP pair with > 6 *discordances because this is expected to eliminate only 1.6% of the SNPs with an error rate of 0.0001 (∼twice the median),* $> 99\%$ *of SNPs with error rate 0.001 and essentially all of the SNPs with error rate of 0.01. (This removes 843 pairs of SNPs.) We also recommend filters for SNPs involved in two annotation problems. (1) There are four duplicate SNPs that are triallelic, in which the minor allele in the dbSNP probe is not the same as the minor allele in the 1000 Genomes probe. (2) Two 1000 Genomes probes have incorrect A/B allele coding relative to the probe sequences. Finally, we recommend filtering out one member of each pair with* $<= 6$ *discordant calls—the one with lowest missing call rate, since these provide redundant information. (This removes one member of each of the remaining 8,920 pairs.)*

Filtering information for the duplicated SNPs and the annotation problems is provided in the file SNP_analysis.csv.

18 Sample exclusion and filtering summary

As discussed in Section 5, genotyping was attempted for a total of 13,129 study samples, of which 12,857 passed CIDR's QC process (Section 2). The subsequent data cleaning QA process identified 12 samples with unresolved identity issues. Therefore, 12,845 study scans will be posted on dbGaP.

We recommend filtering out large chromosomal anomalies associated with error-prone genotypes and whole samples with missing call rate $> 2\%$. *We also recommend filters for specific types of analyses, such as PCA, HWE and association testing as indicated in those sections of this report and summarized in Appendix B. These filters generally include just one scan per subject (unduplicated) and one subject per family (unrelated). PLINK files are provided both with and without chromosome anomalies filtered.*

19 SNP filter summary

Table 1 summarizes SNP failures applied by CIDR prior to data release and a set of additional filters suggested for removing assays of low quality or informativeness. The suggested composite quality filter (for all rows except the final MAF< 0.01 row) is provided as a TRUE/FALSE vector in the "SNP_analysis.csv" file, which also has the individual quality metrics so that the user can apply alternative thresholds. The recommended filters remove 9.90% of the 2,443,179 SNP assays attempted. 25% of these SNPs are monomorphic and therefore completely uninformative.

In addition to the quality filter, we also suggest applying a minor allele frequency filter of at least 1%, as the accuracy of genotype calls decreases for SNPs with MAF < 0.01 (Figure 23d. For illustration, Table 1 provides figures for applying a filter of MAF < 0.01 among unrelated study subjects. The quality and MAF filters combined remove 31.1% of the SNP assays attempted.

Regardless of what filters are applied to association test results, it is highly recommended to view SNP cluster plots for any SNPs of interest.

20 Preliminary association tests

This section has been masked for public release. The complete report is available upon application and approval by dbGaP.

[3] http://www.ncbi.nlm.nih.gov/projects/SNP/

Appendix

A Project participants

University of Michigan
David R. Weir, Jessica Faul, Sharon Kardia, and Jennifer Smith

Center for Inherited Disease Research, Johns Hopkins University
Kim Doheny, Jane Romm, Michelle Zilka, and Tameka Shelford

Genetics Coordinating Center, Department of Biostatistics, University of Washington
Stephanie M. Gogarten, Cathy Laurie, and Bruce Weir

dbGaP-NCBI, National Institutes of Health
Nataliya Sharopova

B Summary of scan selection

A table of chromosome anomalies is provided in the file "chromosome_anomalies.csv." The "filter" column in this table indicates which anomalies are recommended for filtering.

PLINK-format data files are provided with data for all SNPs and one scan per subject. Another PLINK data file is provided with all genotypes zeroed out (i.e. set to missing) for chromosome anomalies > 10Mb with "filter"=TRUE in "chromosome_anomalies.csv." In this filtered PLINK file, all XY (pseudoautosomal) genotypes are set to missing for samples where the entire X chromosome is filtered, and all Y chromosome SNPs for females are set to missing.

The whole-scan filters summarized below are recommended for specific analyses and are given as TRUE/FALSE vectors in the file "Sample_analysis.csv."

- *geno.cntl* designates HapMap controls (0=study, 1=HapMap; in "Sample_annotation.csv.")
- *unrelated* designates one subject per family.
- *pca.eur* designates a set of self-identified Whites with relatively homogeneous ancestry, as those falling within 1 SD of all self-identified Whites for eigenvectors 1 and 2 in the PCA of all unrelated study subjects.
- *hwe.eur* designates scans for the HWE testing of European-ancestry subjects, which are selected as (1) unrelated study subjects, (2) having missing call rate < 2%, (3) self-identified White and (4) belonging to *pca.eur*.
- *pca.afr* designates a set of self-identified African-Americans with relatively homogeneous ancestry, as those falling within 2 SD of all self-identified African-Americans for eigenvector 1 and 1 SD of eigenvector 2 in the PCA of all unrelated study subjects.
- *hwe.afr* designates scans for the HWE testing of African-ancestry subjects, which are selected as (1) unrelated study subjects, (2) having missing call rate < 2%, (3) self-identified African-American and (4) belonging to *pca.afr*.
- *study.02* designates scans for the preliminary association tests as unrelated study subjects with *missing.e2*< 0.02.

For the HWE and association tests, PLINK "keep" files are provided and were created using the above whole-sample filters as described below.

- "hwe_eur_keep.txt" selects samples for which *hwe.eur* is TRUE.
- "hwe_afr_keep.txt" selects samples for which *hwe.afr* is TRUE.
- "assoc_keep.txt" selects scans for which *study.02* is TRUE.

For the association tests, PLINK "extract" files are provided to extract SNPs corresponding to the recommended SNP quality filter and quality plus MAF filter described above.

References

[1] C.C. Laurie et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34:591–602, 2010.

[2] D.A. Peiffer et al. High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Research*, 16:1136–1148, 2006.

[3] L.K. Conlin et al. Mechanisms of mosaicism, chimerism and uniparental disomy identified by single nucleotide polymorphism array analysis. *Human Molecular Genetics*, 19:1263–1275, 2009.

[4] E.S. Venkatraman and A.B. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23:657–663, 2007.

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

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

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

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

[9] W.G. Hill and B.S. Weir. Variation in actual relationship as a consequence of mendelian sampling and linkage. *Genet Res (Camb)*, 93:47–64, 2011.

[10] S. Purcell et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81:559–575, 2007.

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

[12] J. Novembre et al. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.

[13] D.G. Clayton et al. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37:1243–1246, 2005.

[14] A.D. Roses et al. A TOMM40 variable-length polymorphism predicts the age of late-onset Alzheimers disease. *Pharmacogenomics J*, 19:375–384, 2010.

[15] S.C. Johnson et al. The effect of TOMM40 poly-T length on gray matter volume and cognition in middle-aged persons with APOE $\epsilon 3/\epsilon 3$ genotype. *Alzheimers Dement*, 7(4):456–465, 2011.

[16] C. Cruchaga et al. Association and expression analyses with single-nucleotide polymorphisms in TOMM40 in Alzheimer disease. *Arch Neurol*, 68(8):1013–1019, 2011.

[17] M.A. Pericak-Vance et al. Linkage studies in familial Alzheimer disease: Evidence for chromosome 19 linkage. *American Journal of Human Genetics*, 48(6):1034–1050, 1991.

[18] E.H. Corder et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*, 261(5123):921–923, 1993.

[19] E.H. Corder et al. Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. *Nature Genetics*, 7:180–184, 1994.

Table 1: Summary of recommended SNP filters. The number of SNPs lost is given for sequential application of the filters in the order given.

Filter	SNPs lost	SNPs kept
SNP probes		2443179
Intensity-only SNPs	0	2443179
CIDR technical filters	64429	2378750
MAF = 0	60705	2318045
Duplicate SNPs	10162	2307883
Missing call rate >= 2%	89017	2218866
> 4 discordant calls in 423 study duplicates	602	2218264
> 1 Mendelian error	1450	2216814
HWE P-value $< 10^{-4}$ in European or African samples	15441	2201373
Sex difference in allelic frequency >= 0.2	2	2201371
Sex difference in heterozygosity > 0.3	0	2201371
Percent of SNPs lost due to quality filters	9.90%	
MAF < 0.01	518989	1682382
Percent of SNPs lost due to quality and MAF filters	31.1%	

Table 2: Summary of DNA samples and genotyping instances (scans).

	Study	HapMap	Both
DNA samples into genotyping production	13129	411	13540
Failed samples	-272	-0	-272
Scans released by genotyping center	12857	411	13268
Scans failing post-release QC	0	0	0
Scans with unresolved identity issues	12	0	12
Scans to post on dbGaP	12845	411	13256

Table 3: Summary of numbers of scans, subjects and subject characteristics.

	Study	HapMap	Both
Scans to post on dbGaP	12845	411	13256
Subjects	12507	88	12595
Replicated subjects	336	87	423
Families ($N > 1$)	84	25	109
Singletons	12335	13	12348

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

$k2$	$k1$	$k0$	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/avuncular/grandparent-grandchild
0.00	0.25	0.75	0.0625	first cousins
0.00	0.00	1.00	0.0	unrelated

Table 5: Summary of SNP genotyping failures and missingness by chromosome type. A=autosomes, M=mitochondrial, U=unknown position, X=X chromosome, XY=pseudoautosomal, Y=Y chromosome. The row 'SNP technical failures' gives the fraction of SNPs that failed QC at the genotyping center. The row 'missing> 0.05' gives the fraction of SNPs that passed QC at the genotyping center and that have a missing call rate (*missing.n1*) > 0.05.

	A	M	U	X	XY	Y
number of probes	2376443	93	7884	56715	463	1581
SNP technical failures	0.0230	0.0000	0.1435	0.1413	0.0324	0.3789
missing > 0.05	0.0086	0.0000	0.0424	0.0014	0.0156	0.0041

Table 6: Probability of observing more than the given number of discordant calls in 423 pairs of duplicate samples, given an assumed error rate. The number of SNPs with a given number of discordant calls is shown in the final column. The recommended threshold for SNP filtering (in **bold**) is > 4 discordant calls.

	Assumed error rate				
# discordant calls	1.0e-05	1.0e-04	1.0e-3	1.0e-2	# SNPs
> 0	0.00842	0.0811	0.571	1.000	120052
> 1	3.55e-05	0.00338	0.208	0.998	23703
> 2	9.96e-08	9.41e-05	0.054	0.990	9683
> 3	2.09e-10	1.97e-06	0.0108	0.969	5039
> 4	**3.50e-13**	**3.29e-08**	**0.00176**	**0.923**	**2896**
> 5	5.37e-16	4.57e-10	0.000241	0.845	1718
> 6	4.99e-17	5.44e-12	2.82e-05	0.735	1069
> 7	4.93e-17	5.65e-14	2.90e-06	0.603	626

Table 7: Probability of observing more than the given number of discordant calls in 9769 pairs of duplicate SNPs, given an assumed error rate. The number of SNP pairs with a given number of discordant calls is shown in the final column. The recommended threshold for SNP filtering (in **bold**) is > 6 discordant calls.

	Assumed error rate				
# discordant calls	1.0e-05	1.0e-04	1.0e-3	1.0e-2	# SNPs
> 0	2.265549e-01	0.923393	1.000000	1.000000	5307
> 1	2.785442e-02	0.726564	1.000000	1.000000	3285
> 2	2.333010e-03	0.473726	1.000000	1.000000	2182
> 3	1.478388e-04	0.257219	1.000000	1.000000	1641
> 4	7.527209e-06	0.118183	1.000000	1.000000	1267
> 5	3.201551e-07	0.046759	0.999999	1.000000	1016
> 6	**1.168936e-08**	**0.016186**	**0.999996**	**1.000000**	**843**
> 7	3.738097e-10	0.004969	0.999986	1.000000	733

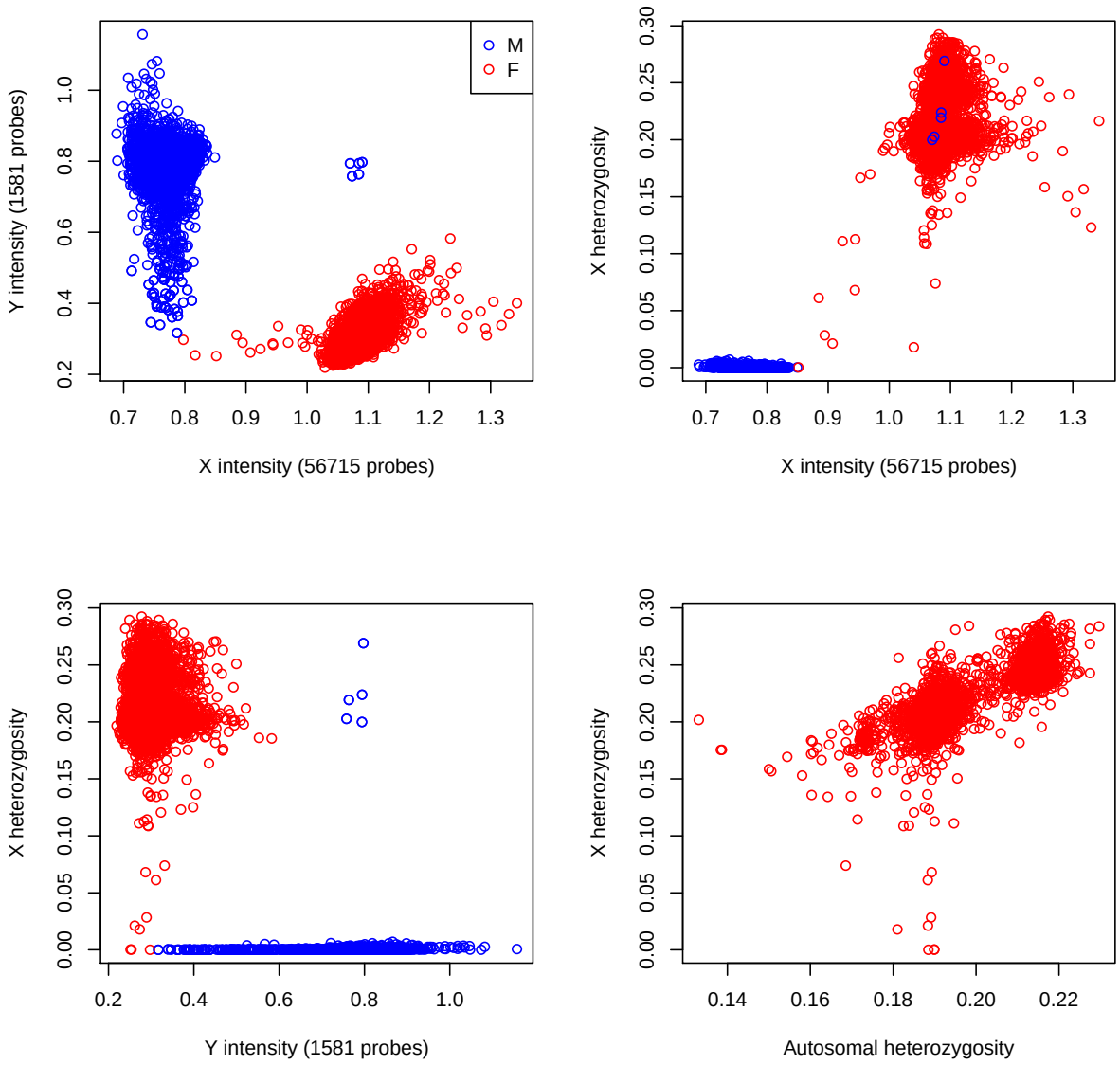


Figure 1: The X and Y intensities are calculated for each sample as the mean of the sum of the normalized intensities of the two alleles for each probe on those chromosomes. Sample sizes are given in the axis labels. X heterozygosity is the fraction of heterozygous calls out of all non-missing genotype calls on the X chromosome for each sample.

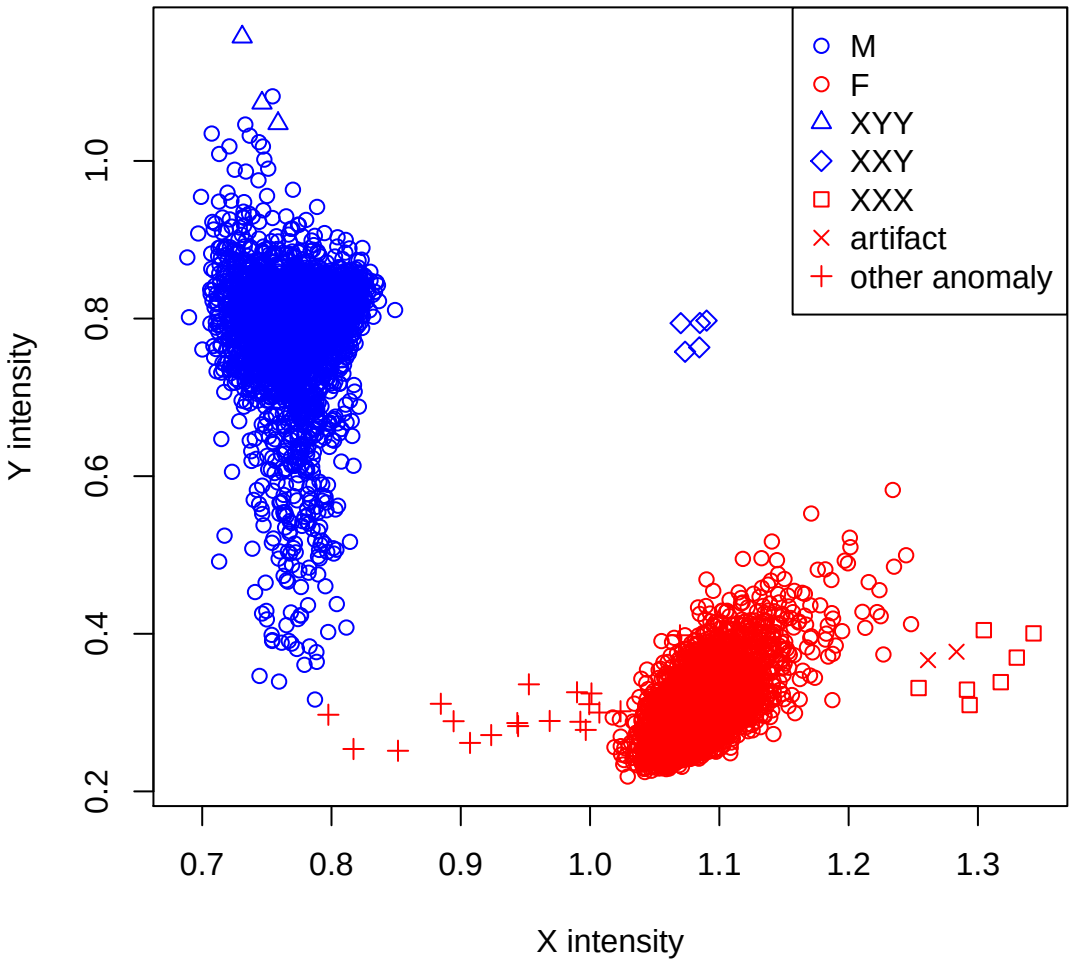


Figure 2: X and Y intensities as in Figure 1, with sex chromosome anomalies annotated.

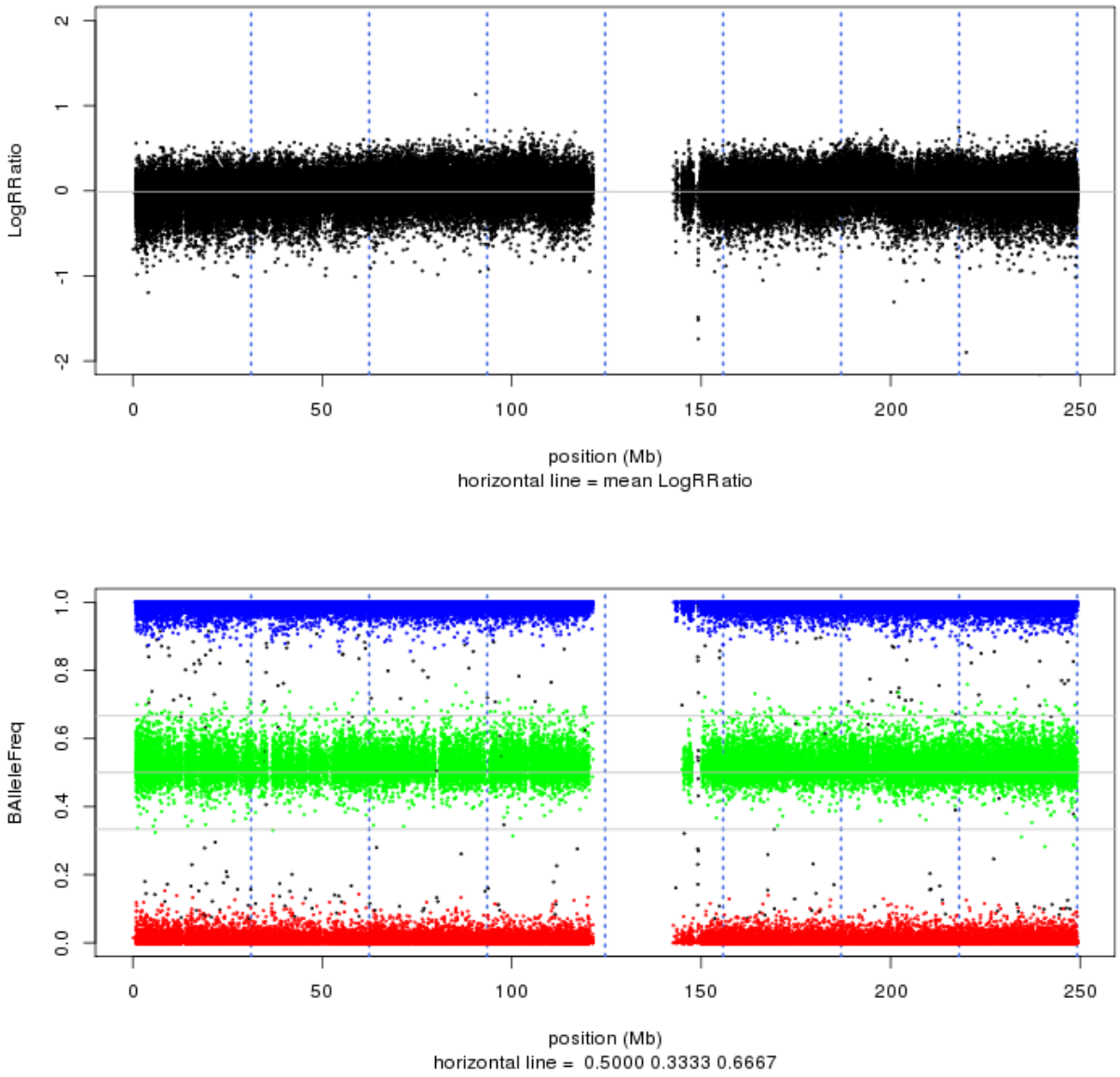


Figure 3: LRR and BAF plots for chromosome 1 in Sample A. This chromosome shows a normal pattern. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing).

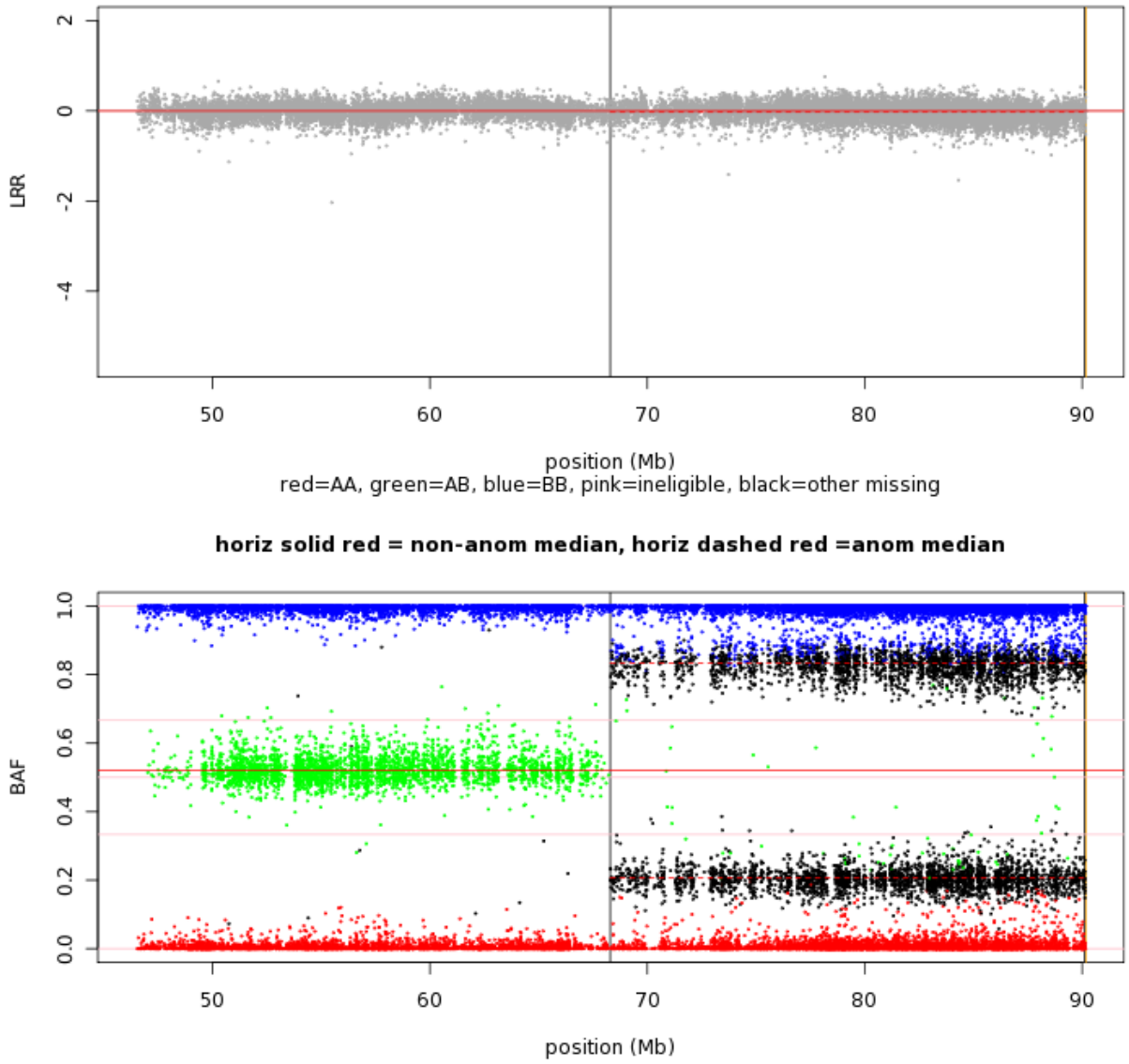


Figure 4: LRR and BAF plots for chromosome 16 in Sample A. This chromosome shows a split in the heterozygous band wide enough to cause genotyping errors, with some heterozygotes called as homozygotes. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing). The horizontal solid red line in both plots is the median value of non-anomalous regions of the autosomes, while the horizontal dashed red line is the median value within the anomaly.

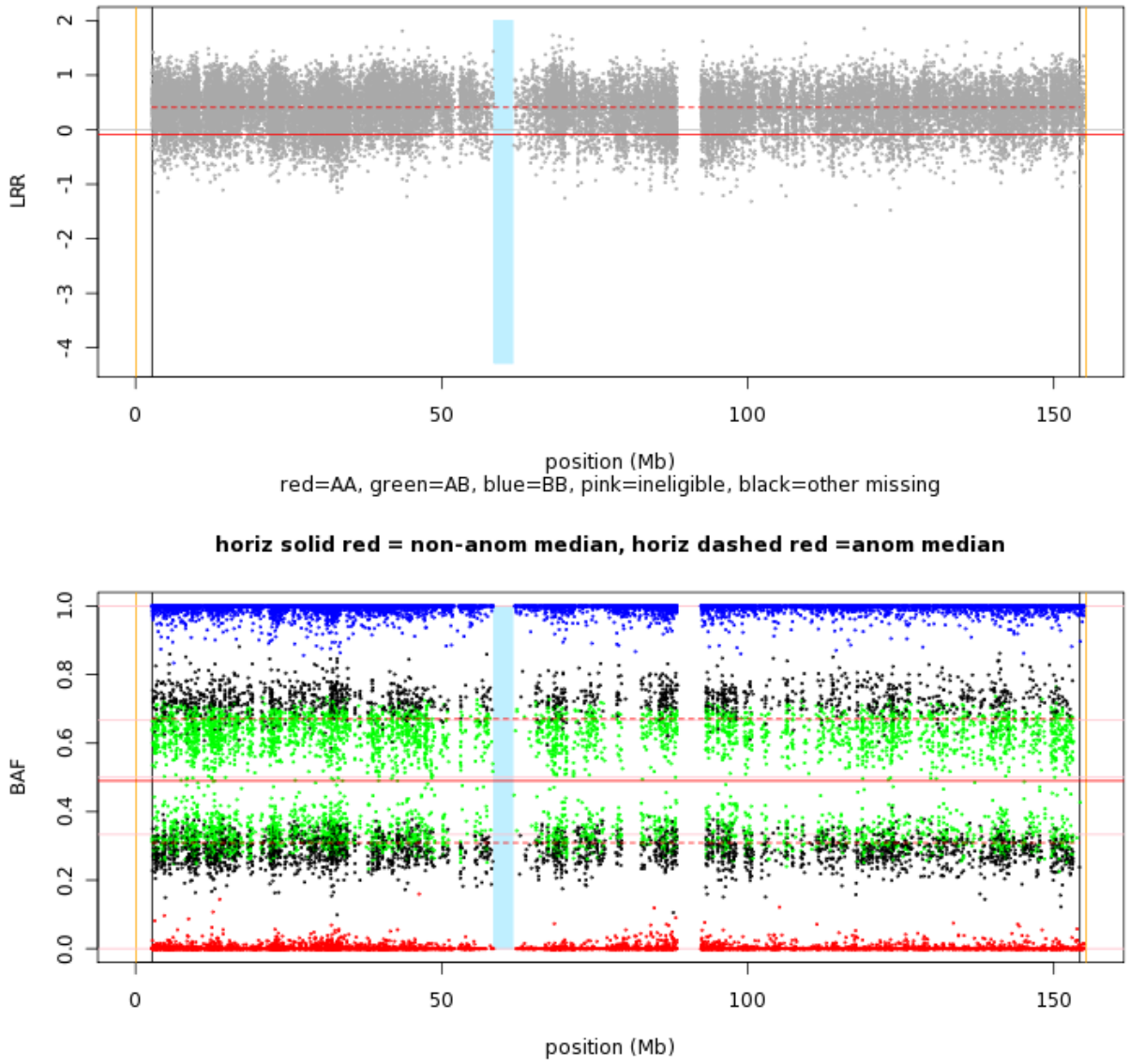


Figure 5: LRR and BAF plots for chromosome X in Female B. This chromosome shows a pattern consistent with trisomy, i.e., a split in the heterozygous band with positions at about 1/3 and 2/3, and elevated LRR. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing). The horizontal solid red line in both plots is the median value of non-anomalous regions of the autosomes, while the horizontal dashed red line is the median value within the anomaly.

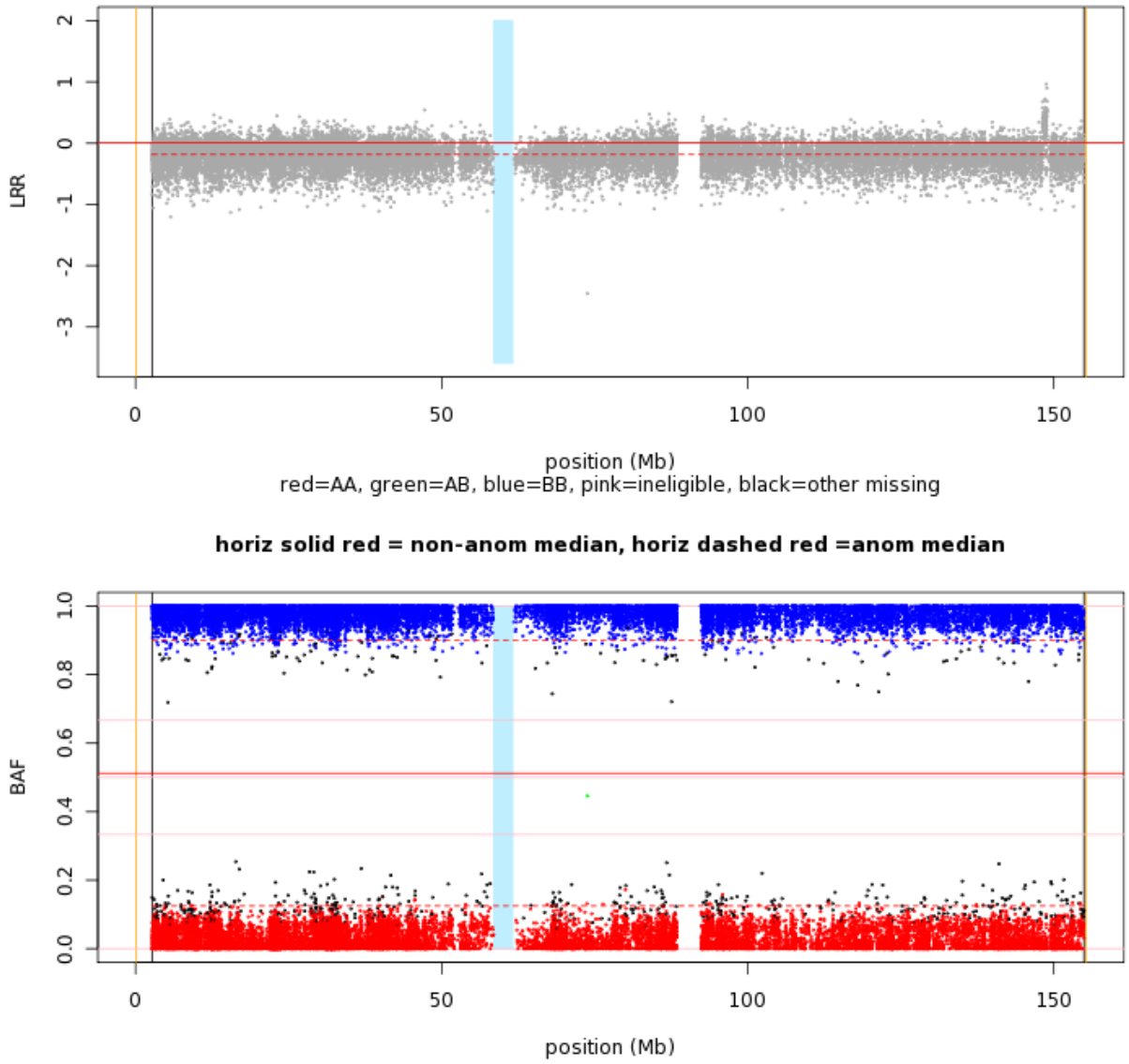


Figure 6: LRR and BAF plots for chromosome X in Female C. This chromosome shows an XX/X0 mosaic in which the heterozygous band has merged with the homozygous bands. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing). The horizontal solid red line in both plots is the median value of non-anomalous regions of the autosomes, while the horizontal dashed red line is the median value within the anomaly.

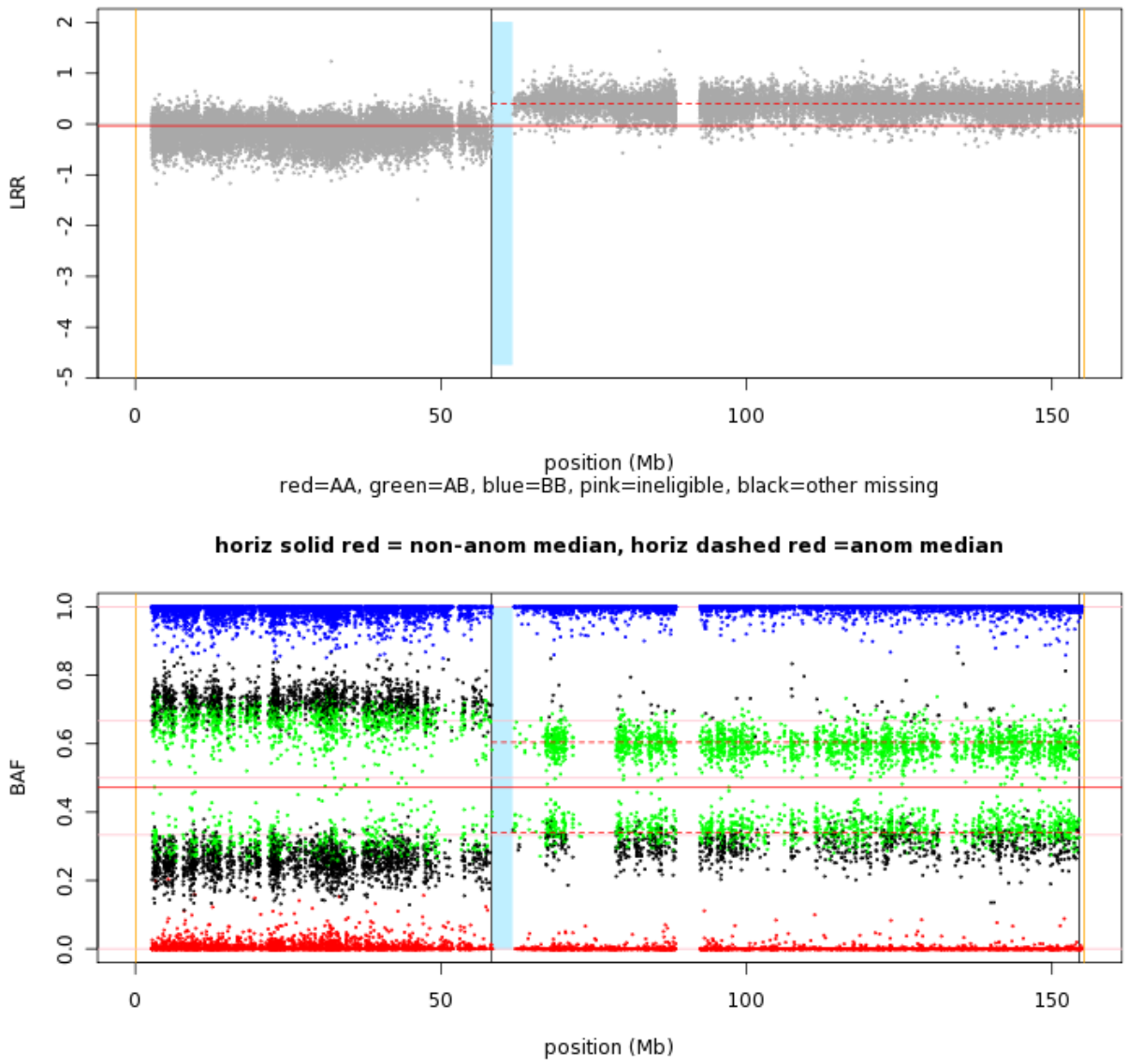


Figure 7: LRR and BAF plots for chromosome X in Female D. This chromosome shows two separate anomalies. Only the anomaly on the left is wide enough to cause genotyping errors; the anomaly on the right will not be filtered. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing). The horizontal solid red line in both plots is the median value of non-anomalous regions of the autosomes, while the horizontal dashed red line is the median value within the anomaly.

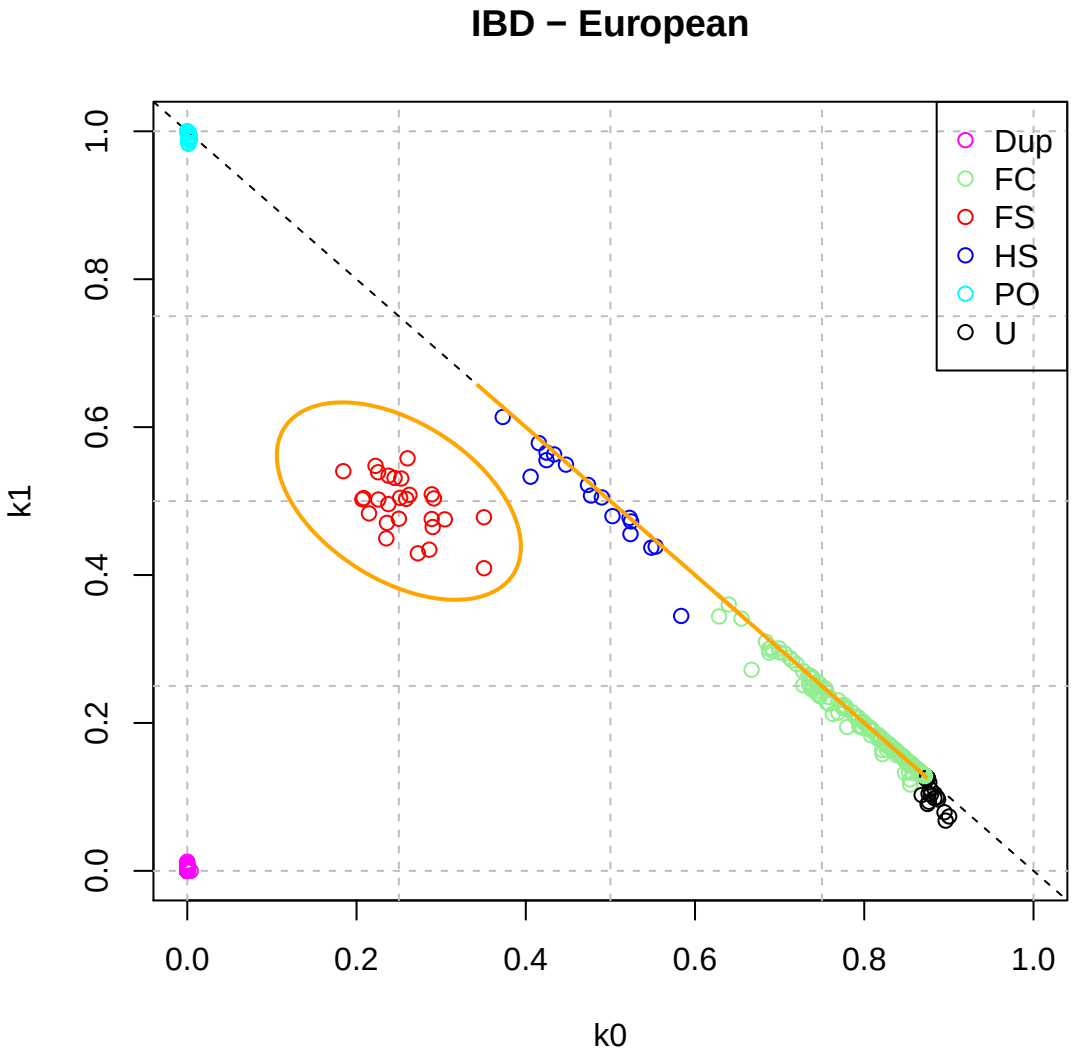


Figure 8: IBD coefficients to estimate relatedness. Each point represents a pair of samples. This plot shows 672 study pairs of PCA-defined European participants with an estimated $KC > 1/32$, color-coded by observed relationships determined by 99.6% prediction ellipse/intervals, as described in the text. The diagonal line is $k0 + k1 = 1$ and the 99.6% prediction ellipse (for full siblings) and intervals (for half-sibling-like and first cousins, which merge together) are shown in orange. In the legend, "Dup" = duplicates, "FC" = first cousins, "FS" = full siblings, "HS"=half-sibling-like relationships, "PO" = parent-offspring, and "U" = unrelated samples.

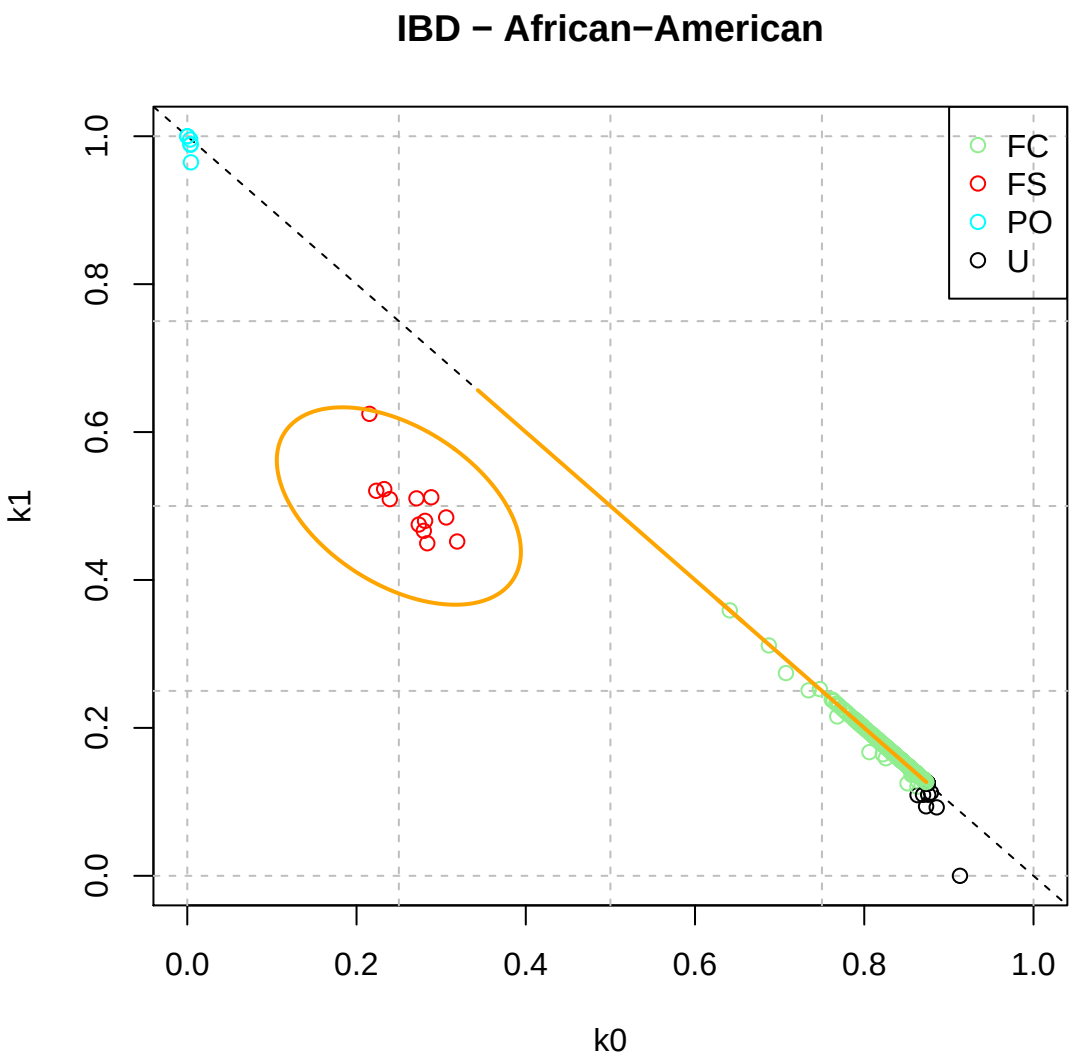


Figure 9: IBD coefficients to estimate relatedness. Each point represents a pair of samples. This plot shows 1077 study pairs of PCA-defined African-American participants with an estimated $KC > 1/32$, color-coded by observed relationships determined by 99.6% prediction ellipse/intervals, as described in the text. The diagonal line is $k0 + k1 = 1$ and the 99.6% prediction ellipse (for full siblings) and intervals (for half-sibling-like and first cousins, which merge together) are shown in orange. In the legend, "FC" = first cousins, "FS" = full siblings, "PO" = parent-offspring, and "U" = unrelated samples.

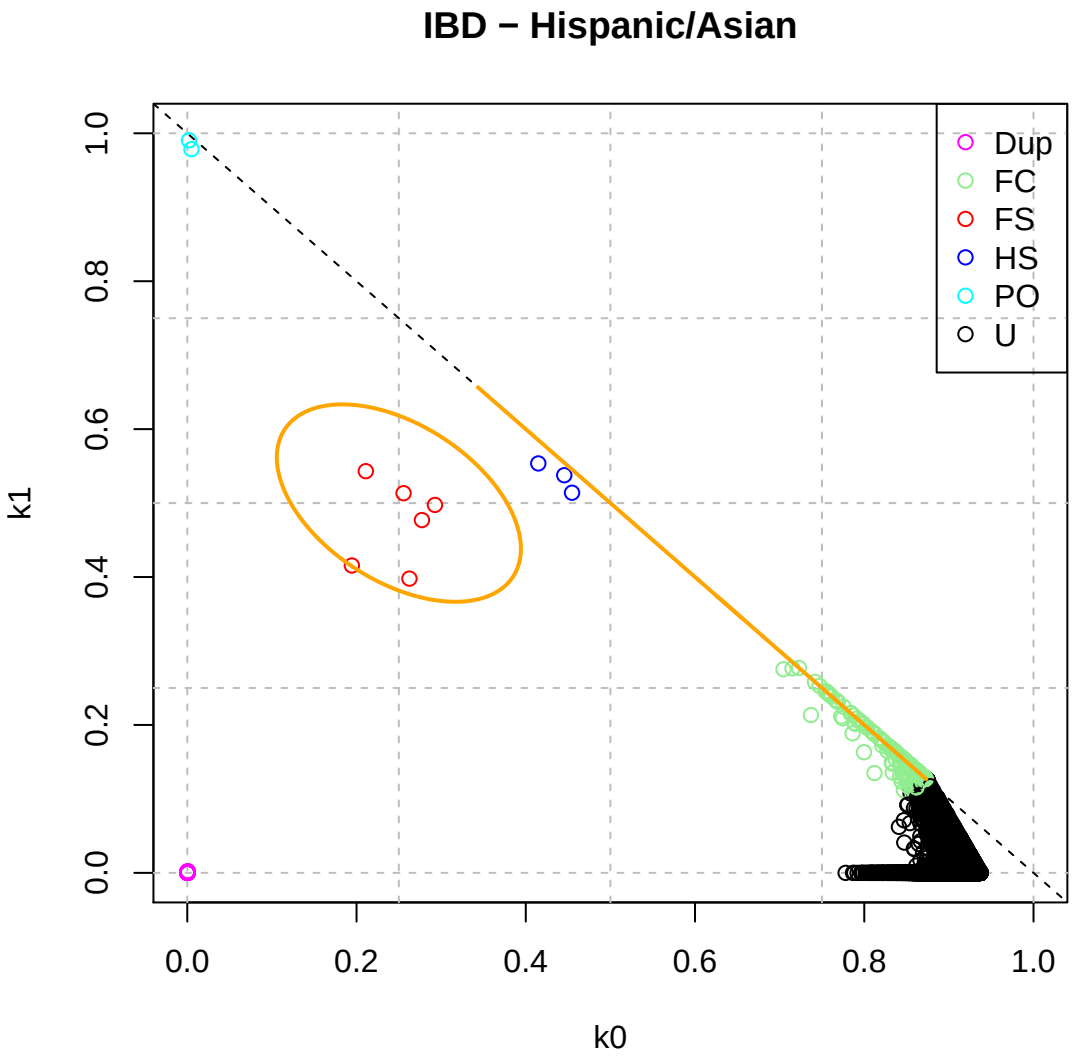


Figure 10: IBD coefficients to estimate relatedness. Each point represents a pair of samples. This plot shows 12303 study pairs of PCA-defined Hispanic/Asian participants with an estimated $KC > 1/32$, color-coded by observed relationships determined by 99.6% prediction ellipse/intervals, as described in the text. The diagonal line is $k0 + k1 = 1$ and the 99.6% prediction ellipse (for full siblings) and intervals (for half-sibling-like and first cousins, which merge together) are shown in orange. In the legend, "Dup" = duplicates, "FC" = first cousins, "FS" = full siblings, "HS"=half-sibling-like relationships, "PO" = parent-offspring, and "U" = unrelated samples.

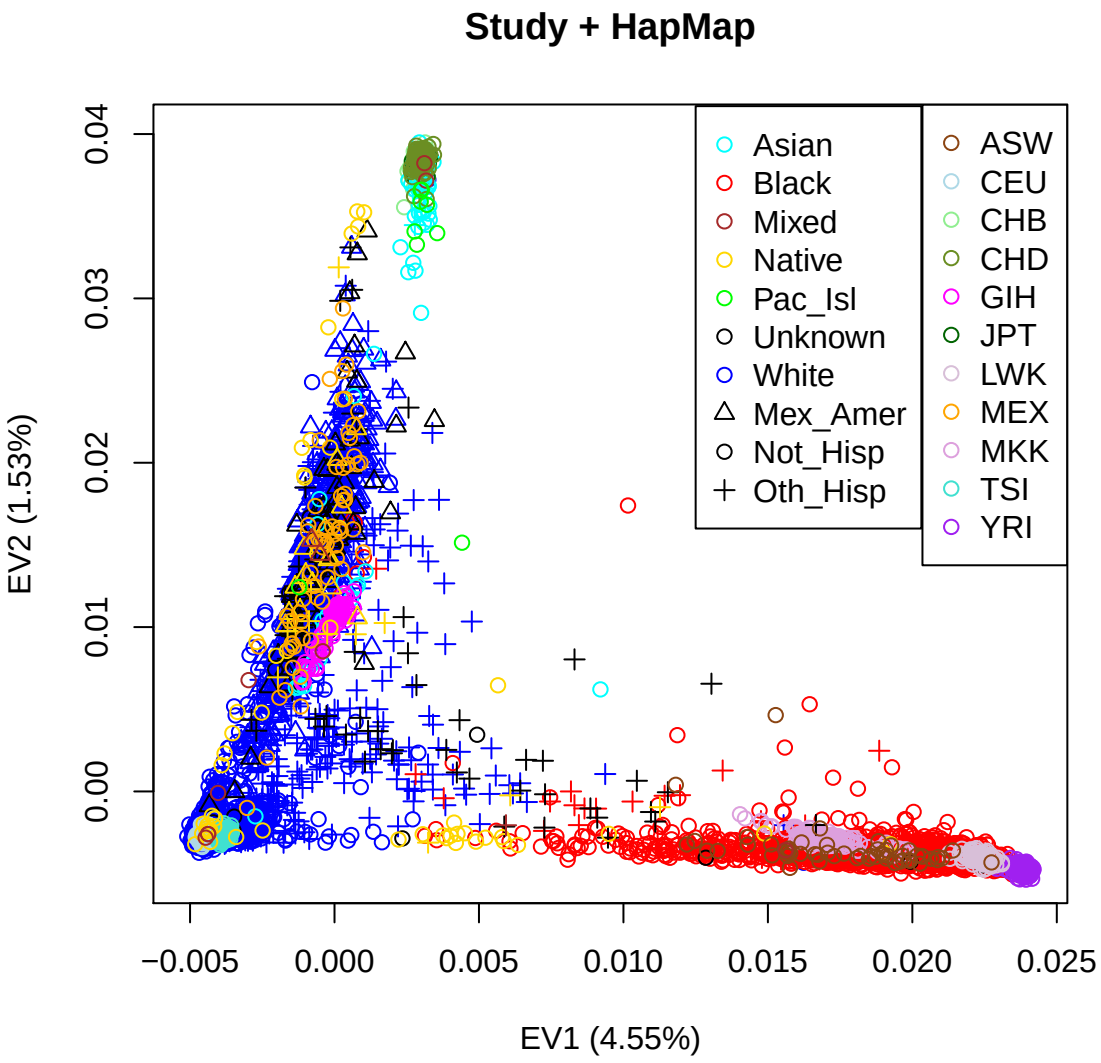
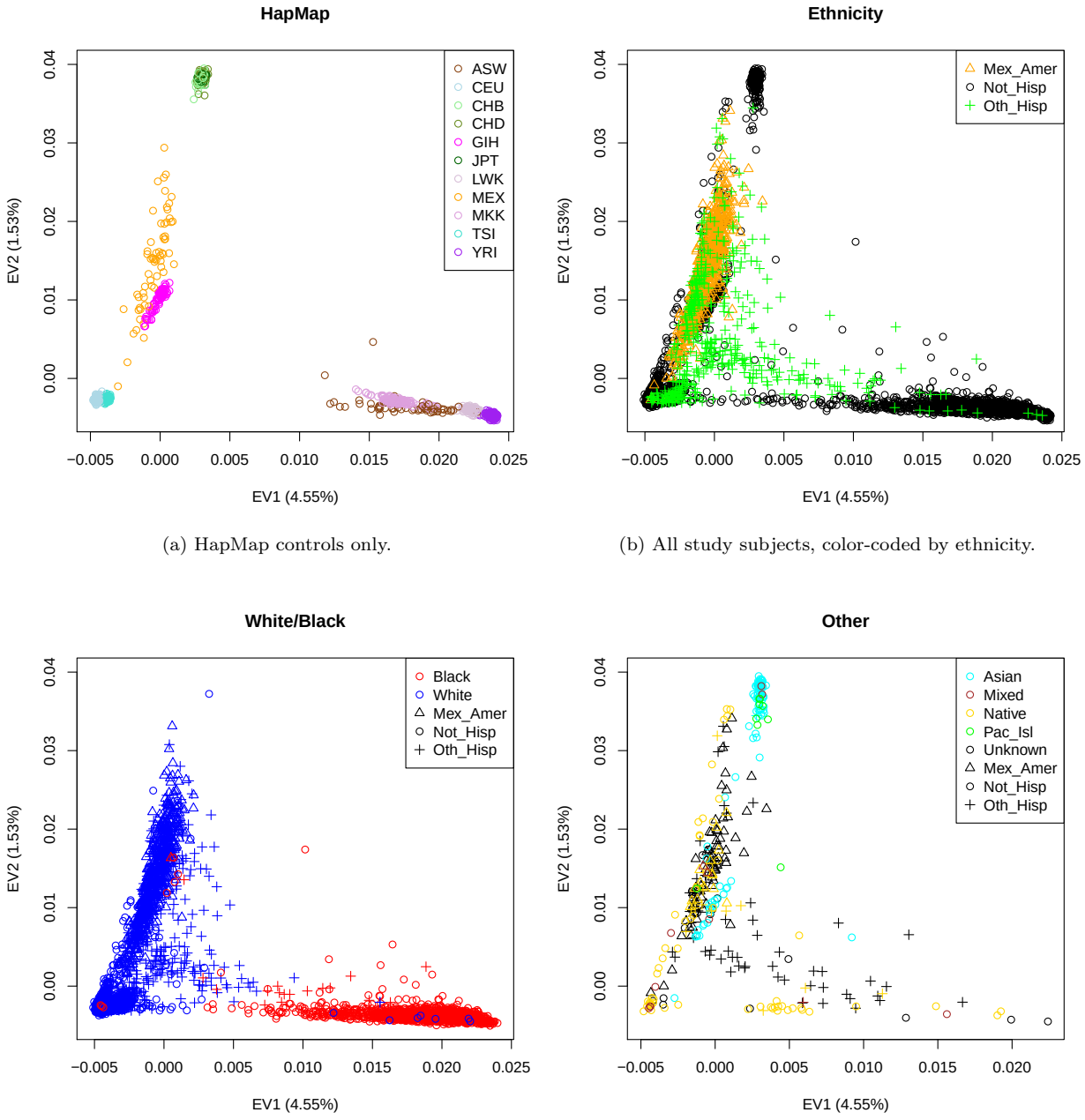


Figure 11: Principal component analysis of 12,507 study subjects with 1230 HapMap controls. Color-coding is according to self-identified race, while symbol denotes ethnicity (Hispanic or not). Axis labels indicate the percentage of variance explained by each eigenvector.



(a) HapMap controls only.

(b) All study subjects, color-coded by ethnicity.

(c) Self-identified white and black study subjects only.

(d) Study subjects with races other than white or black.

Figure 12: PCA results from Figure 11, broken down by subgroup for clarity.

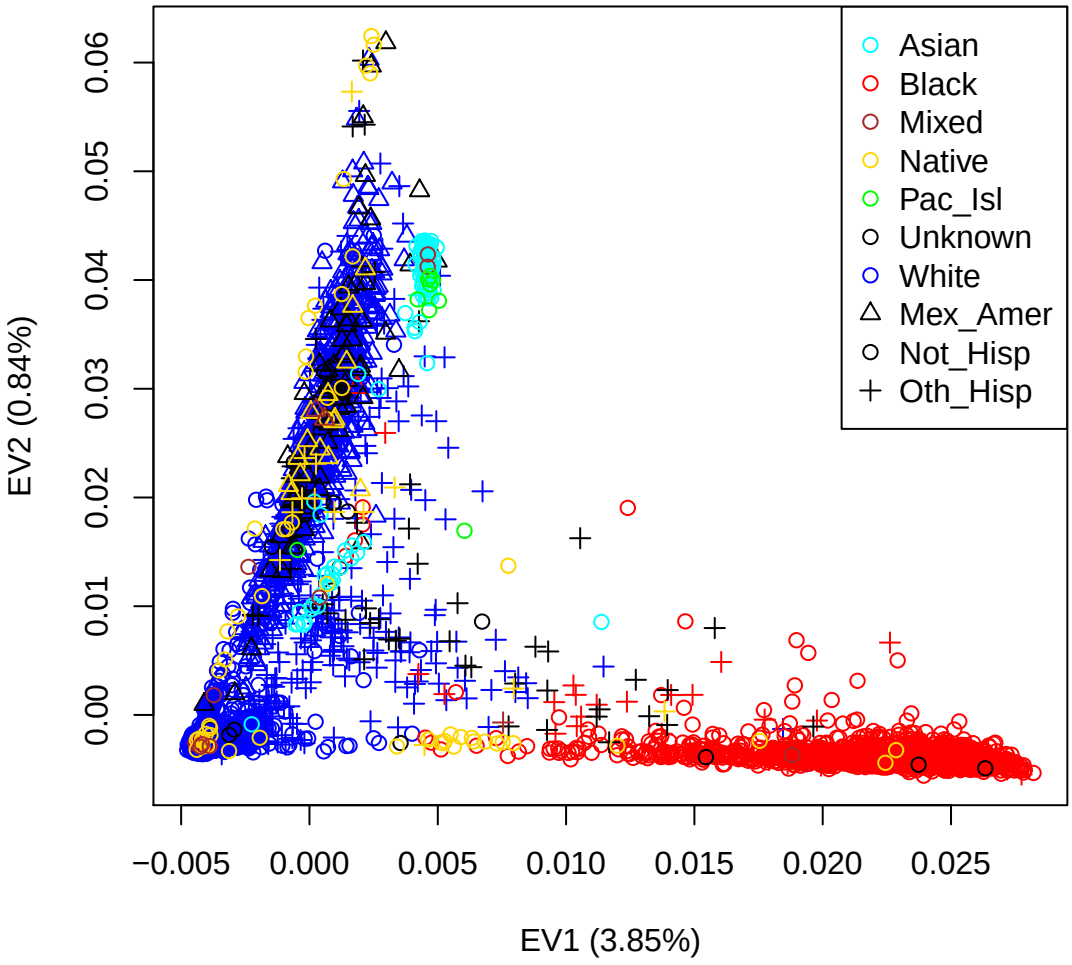


Figure 13: Principal component analysis of 12,419 unrelated study subjects without HapMap controls. Color-coding is according to self-identified race, while symbol denotes ethnicity (Hispanic or not). Axis labels indicate the percentage of variance explained by each eigenvector.

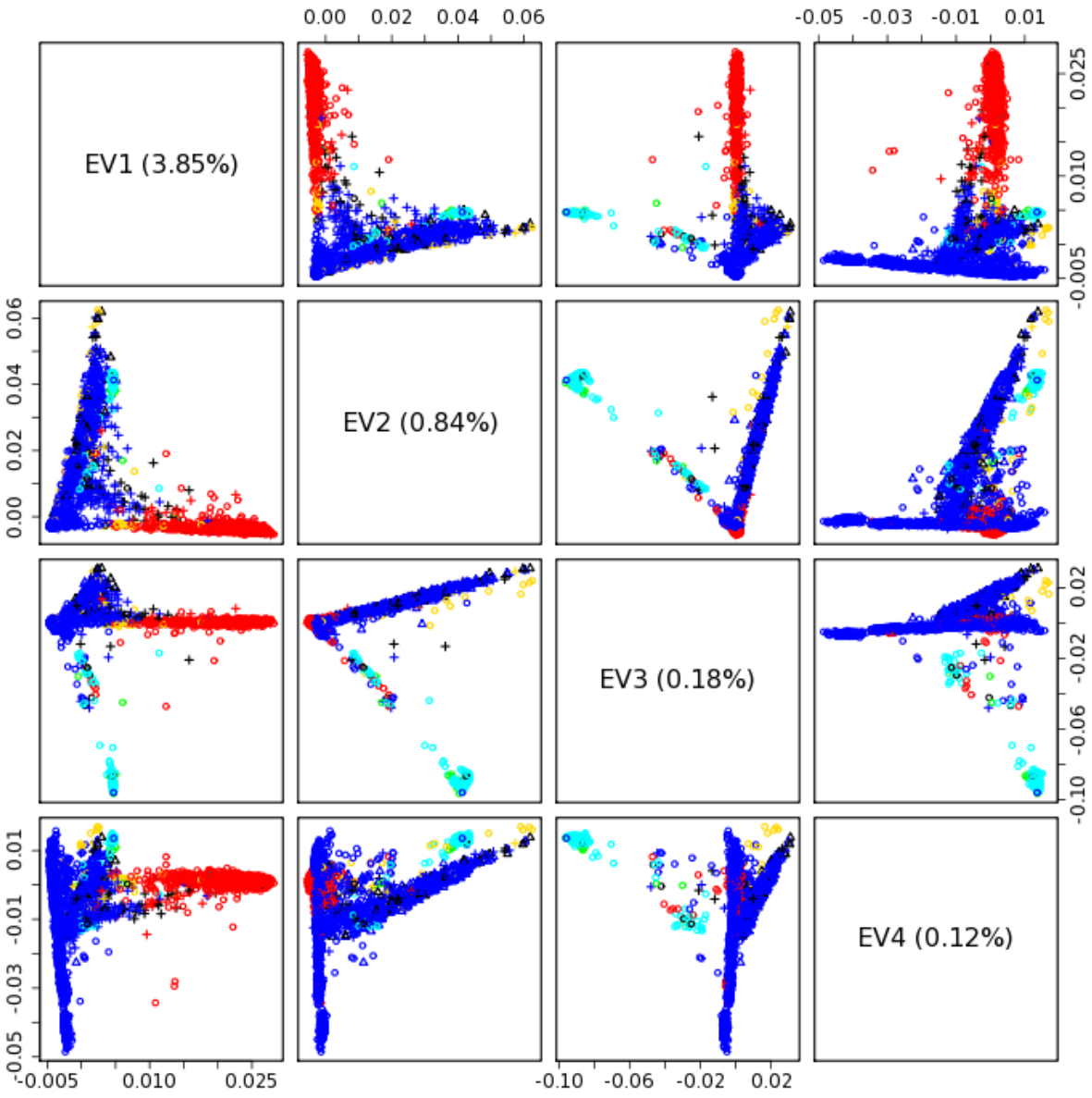


Figure 14: Principal component analysis of 12,419 unrelated study subjects without HapMap controls. Pairwise plots of the first four eigenvectors are shown and the percent of variance accounted for by each is given in parentheses along the diagonal. Color-coding and symbols are as in as in Figure 13.

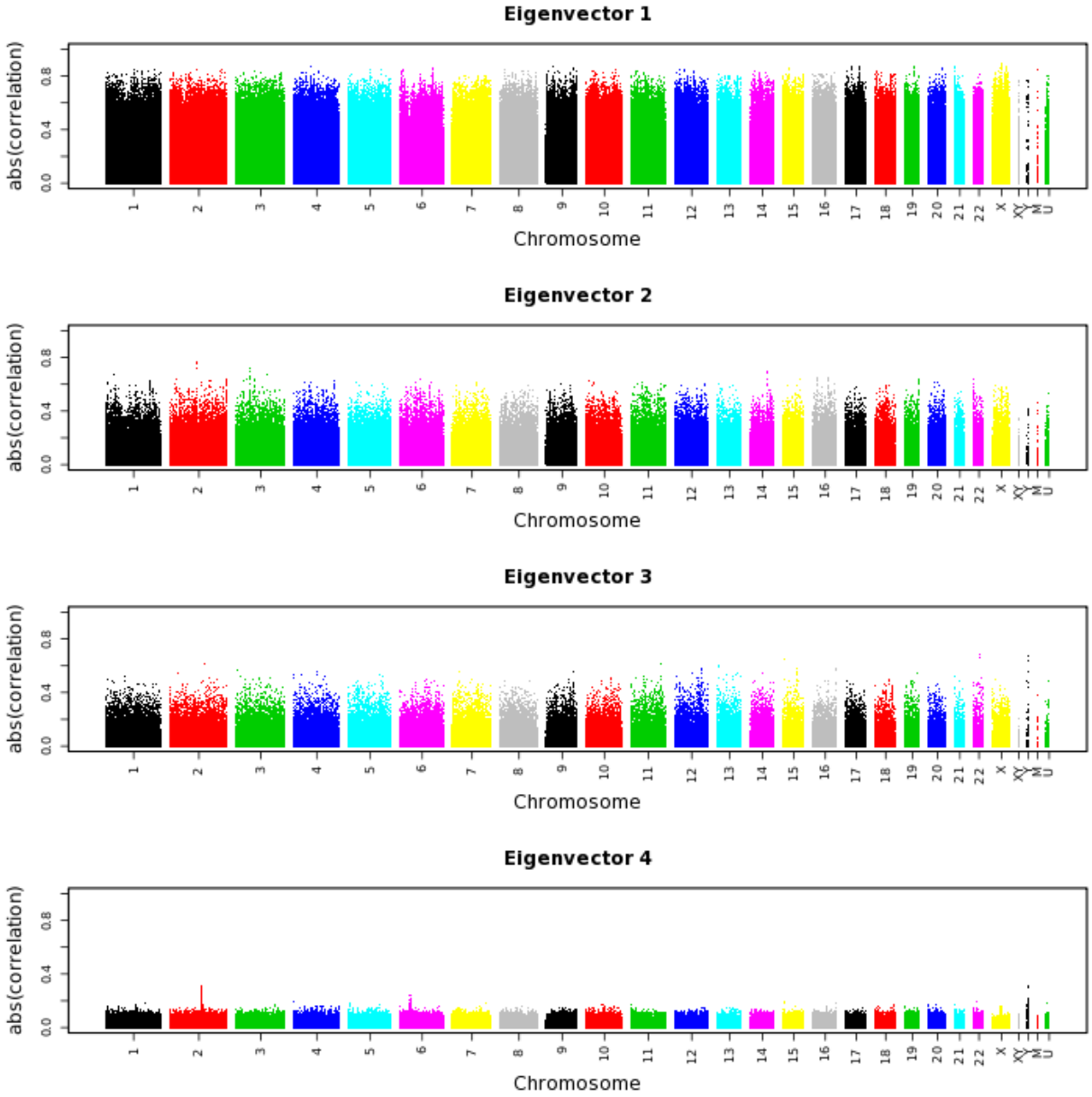


Figure 15: SNP position versus correlation between SNP genotype (0, 1 or 2) and each of the first 8 eigenvectors. These eigenvectors are from the PCA of all unrelated study subjects.

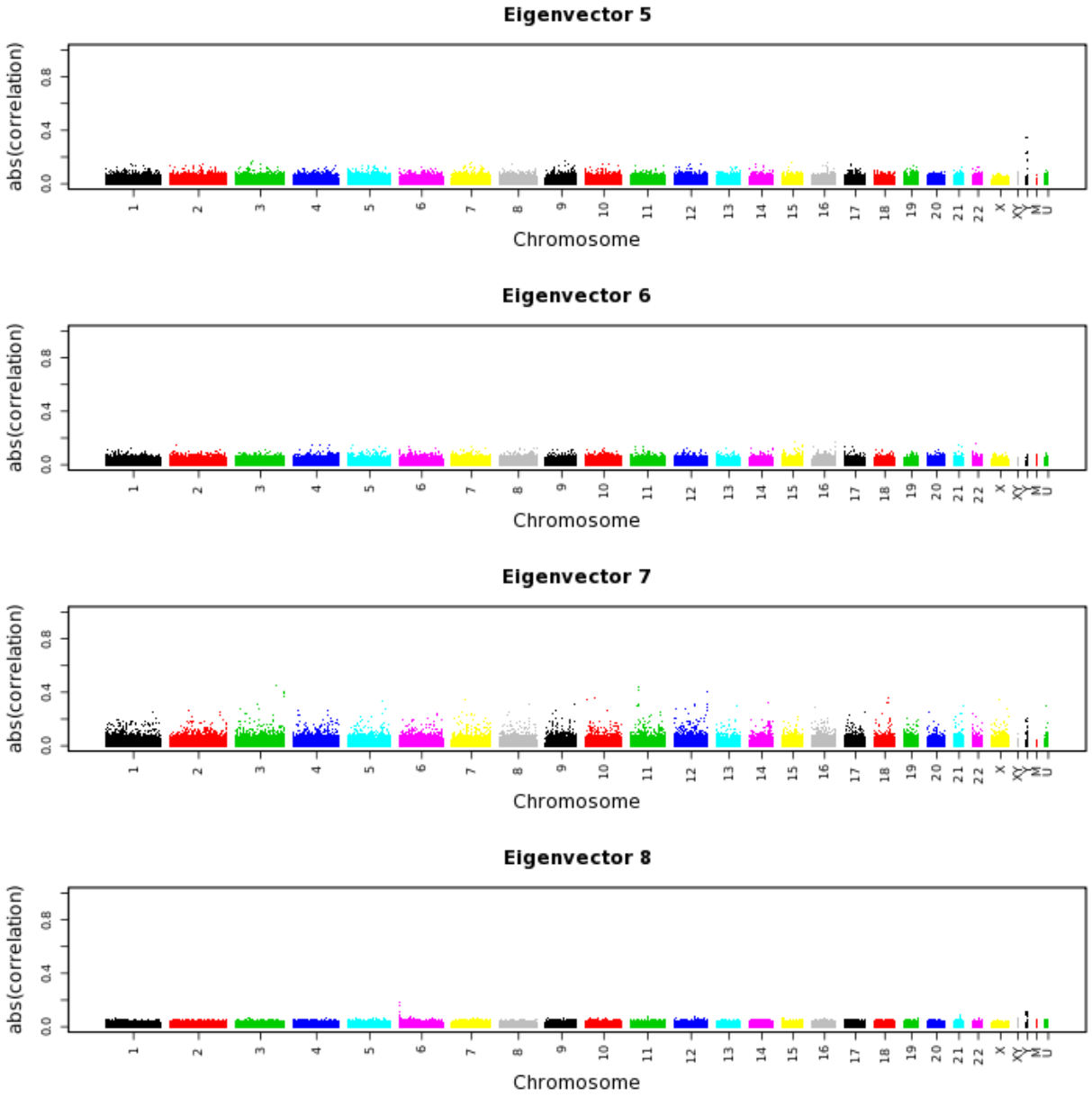


Figure 15: Continued.

Table 8: P-values for linear regression of total recall score on the first eight eigenvectors.

Eigenvector	P-value	signif
1	< 2e-16	***
2	< 2e-16	***
3	6e-05	***
4	0.00077	***
5	0.47585	
6	0.00155	**
7	0.39655	
8	0.39551	

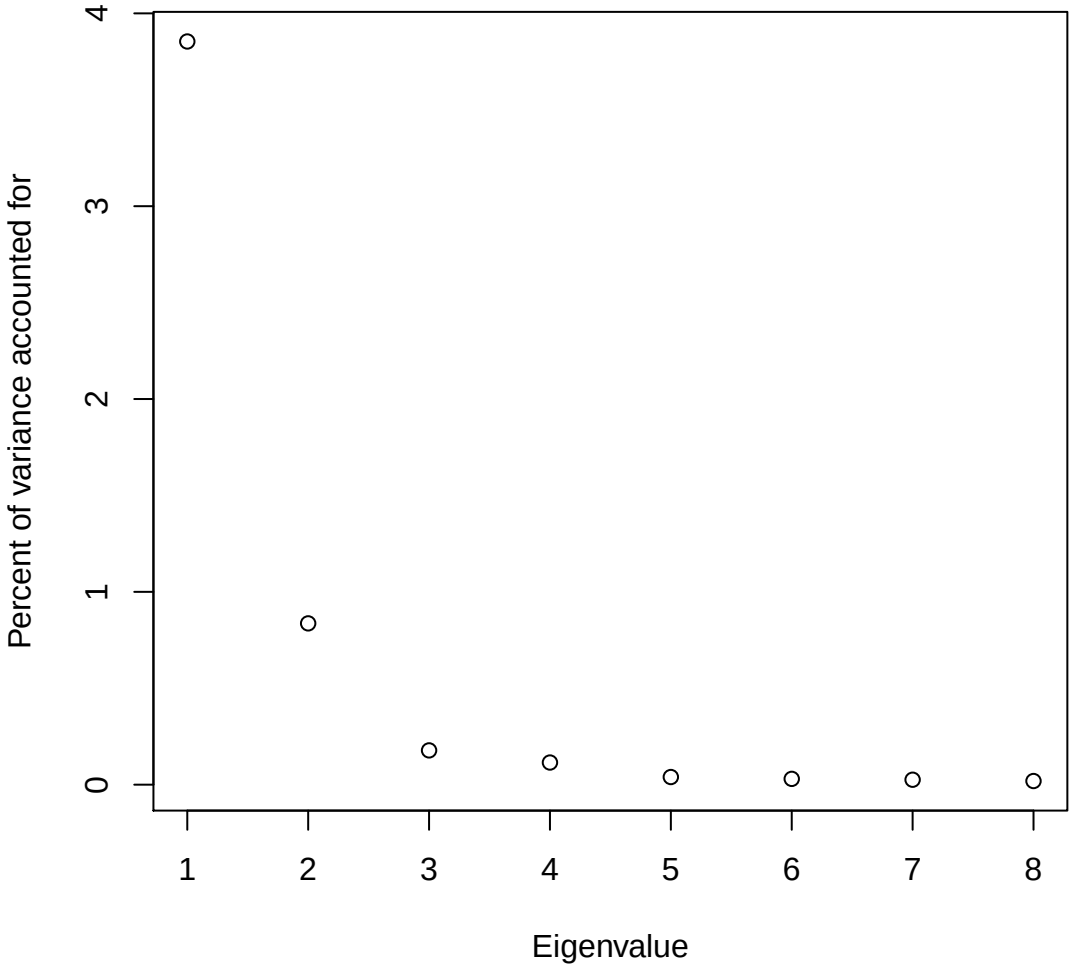


Figure 16: Scree plot for PCA shown in Figure 13.

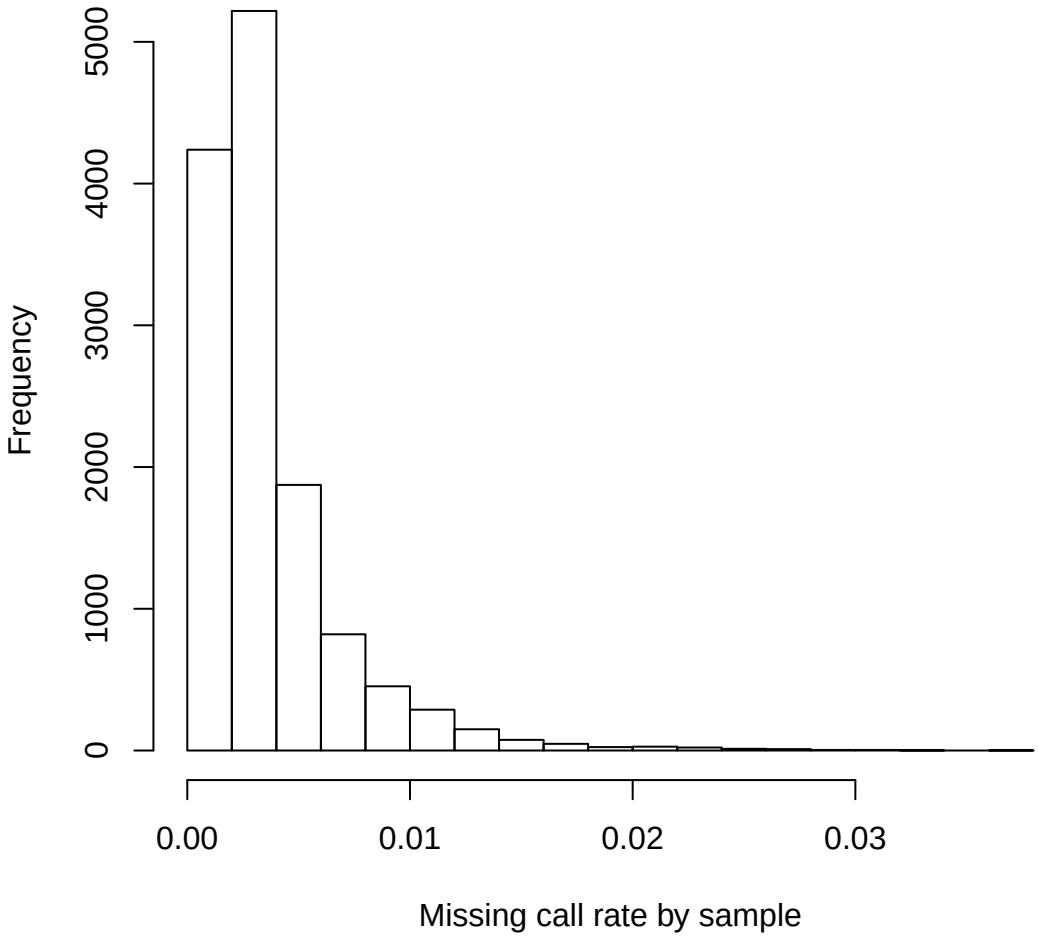
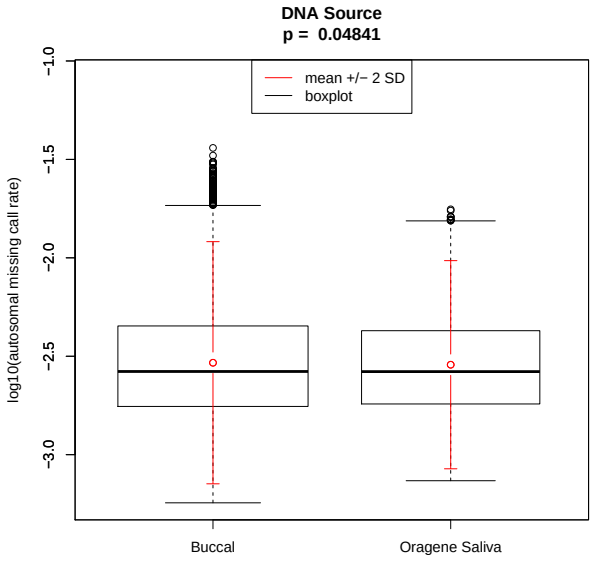
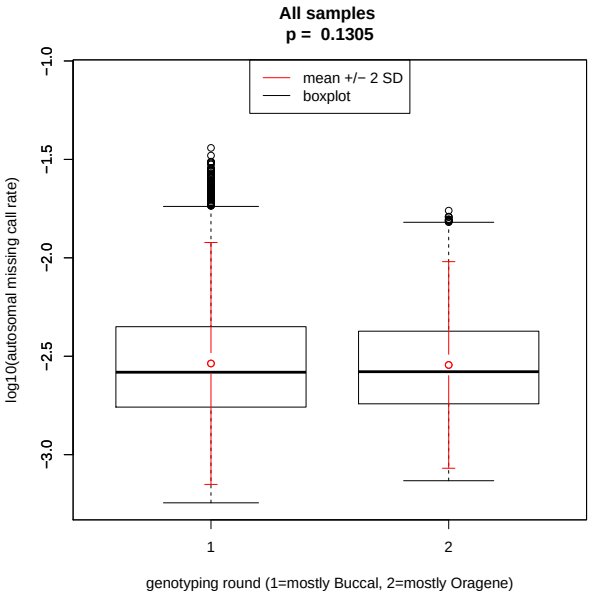


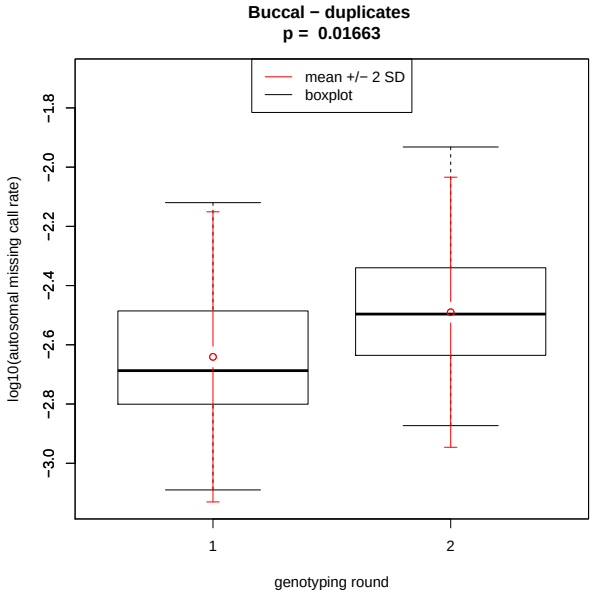
Figure 17: Histogram of the missing call rate per sample (*missing.e1*).



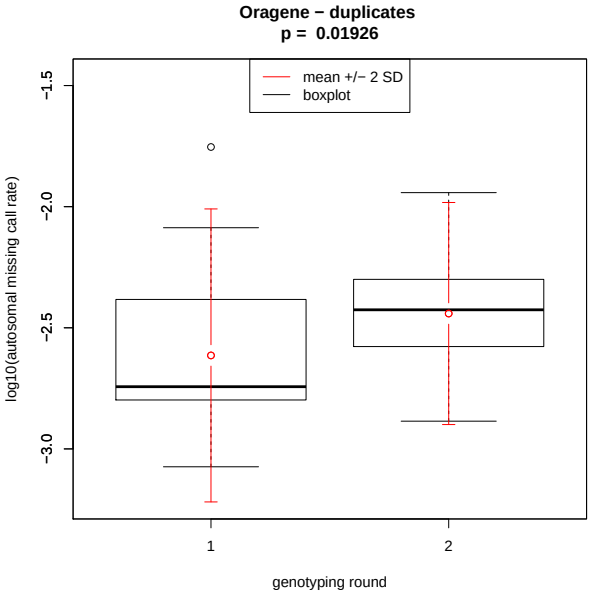
(a) All samples by DNA source.



(b) All samples by genotyping round.



(c) Buccal duplicates run in both rounds only.



(d) Oragene duplicates run in both rounds only.

Figure 18: Comparison of missing call rates by sample between Buccal and Oragene samples, which were genotyped separately but called together.

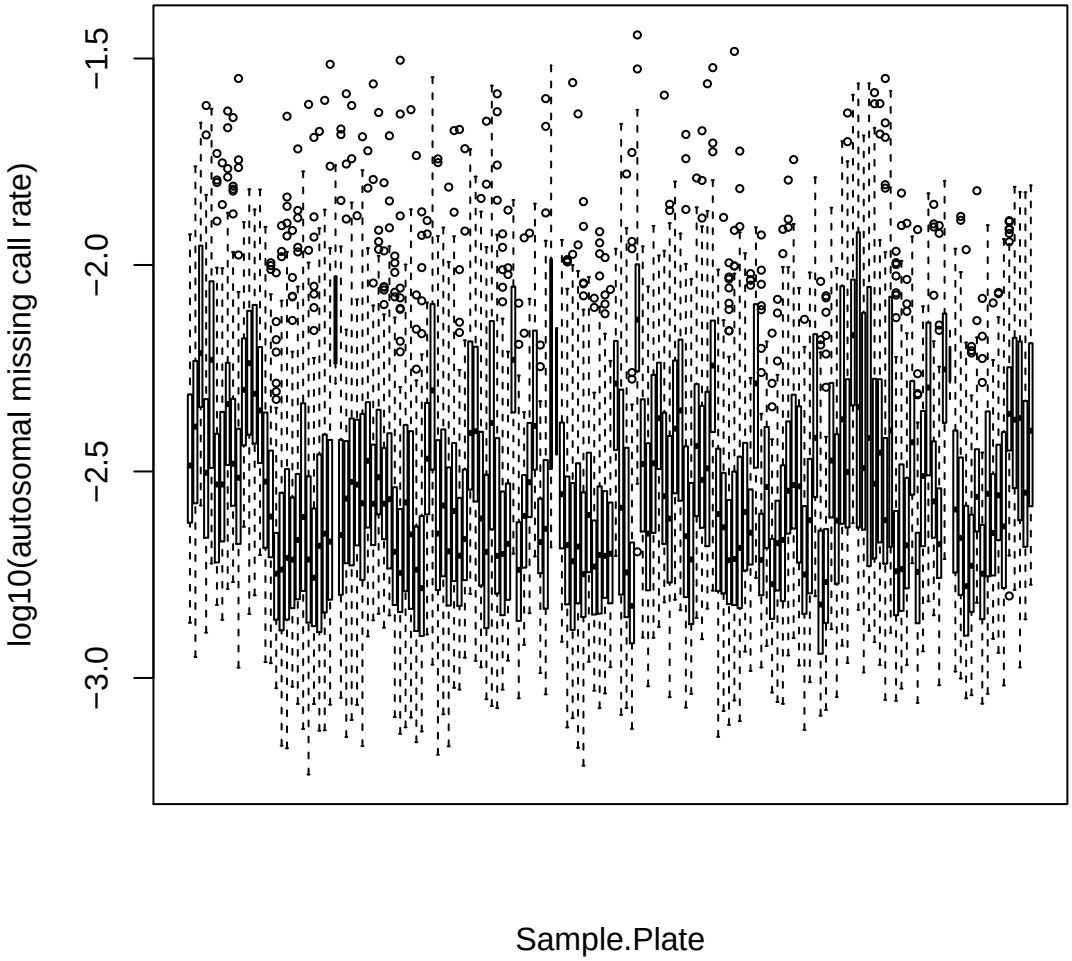


Figure 19: Boxplot of missing call rate for study samples categorized by plate.

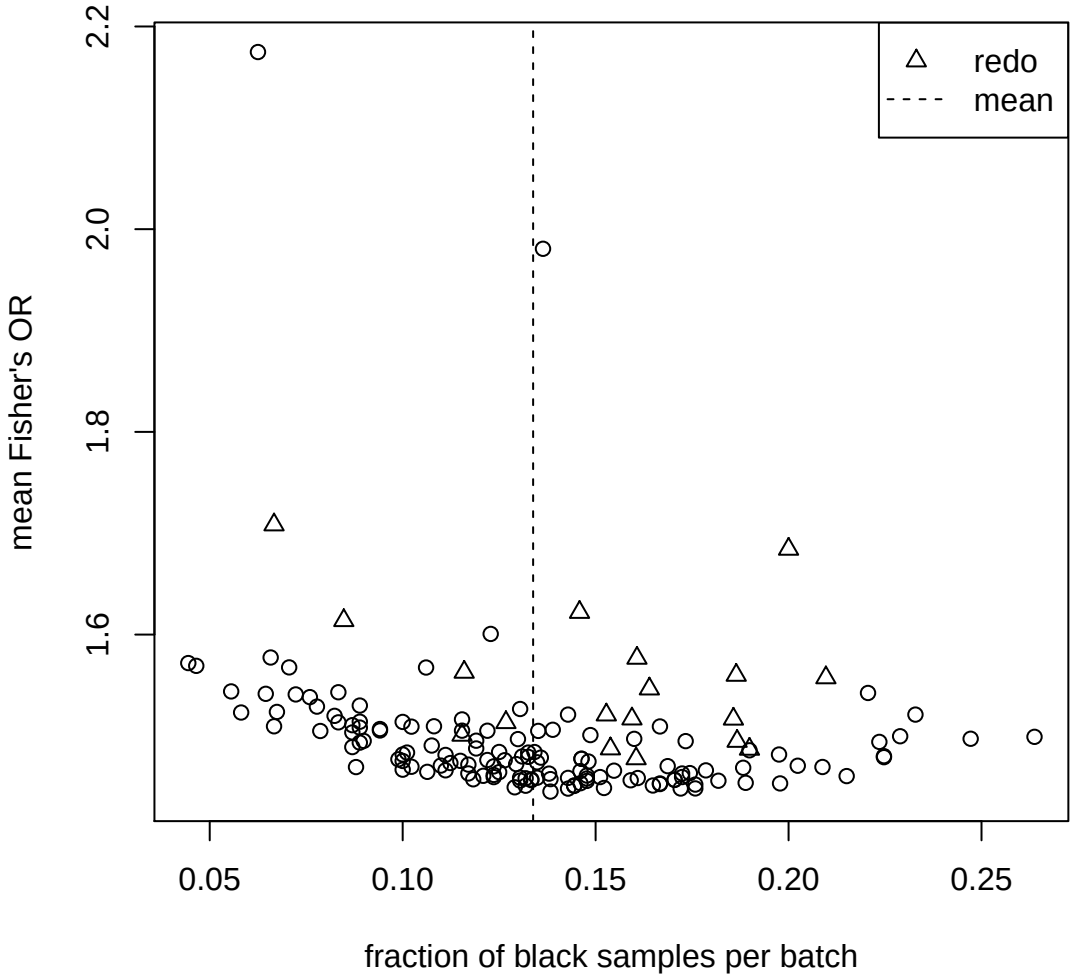


Figure 20: Mean odds ratio from Fisher's exact test of allele frequency plotted against fraction of self-identified black samples per plate. The dashed vertical line is the mean over all plates.

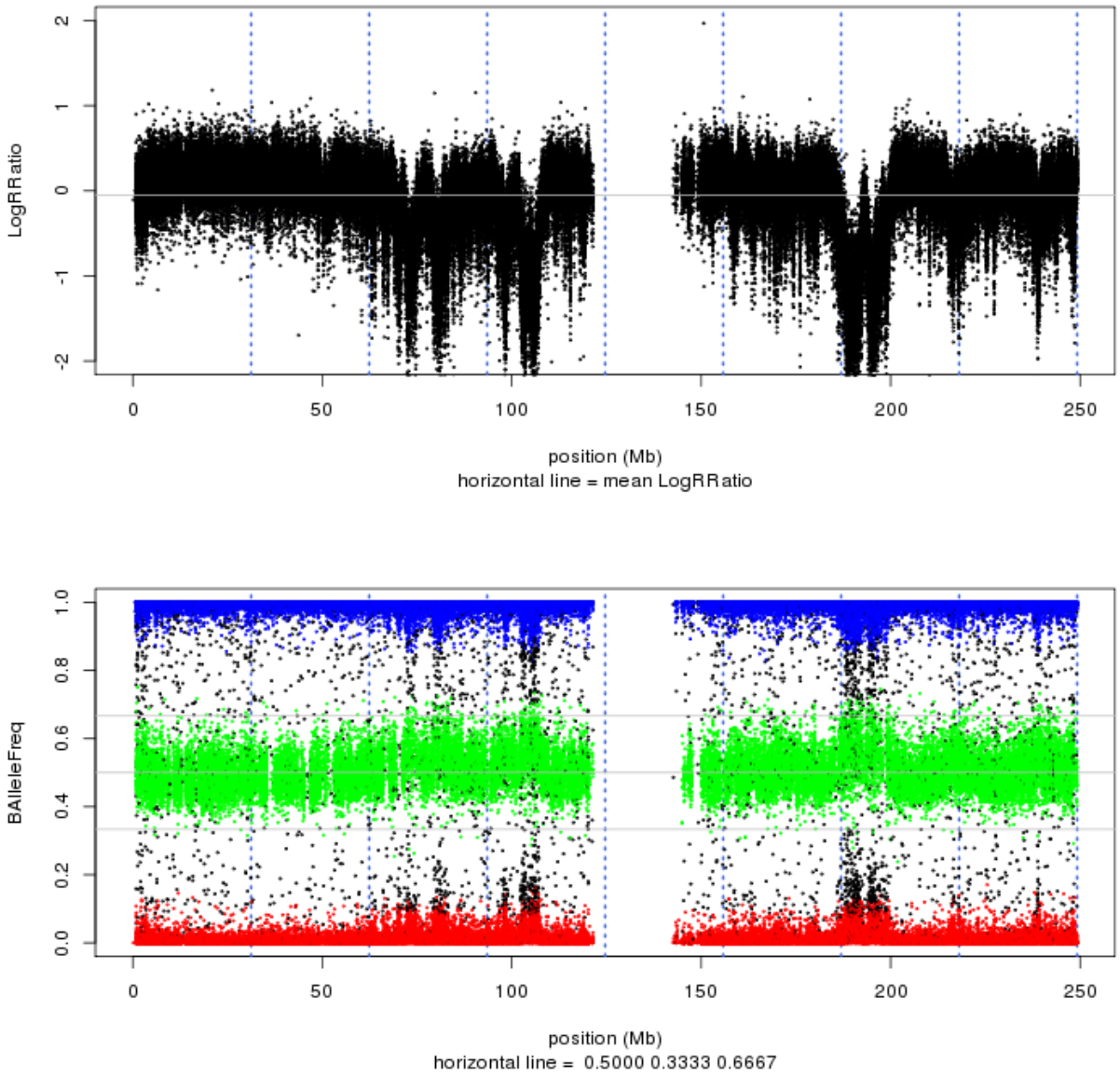


Figure 21: LRR and BAF plots for chromosome 1 in Sample E. This chromosome shows an extreme example of the LRR waves common to low-concentration Oragene samples. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing).

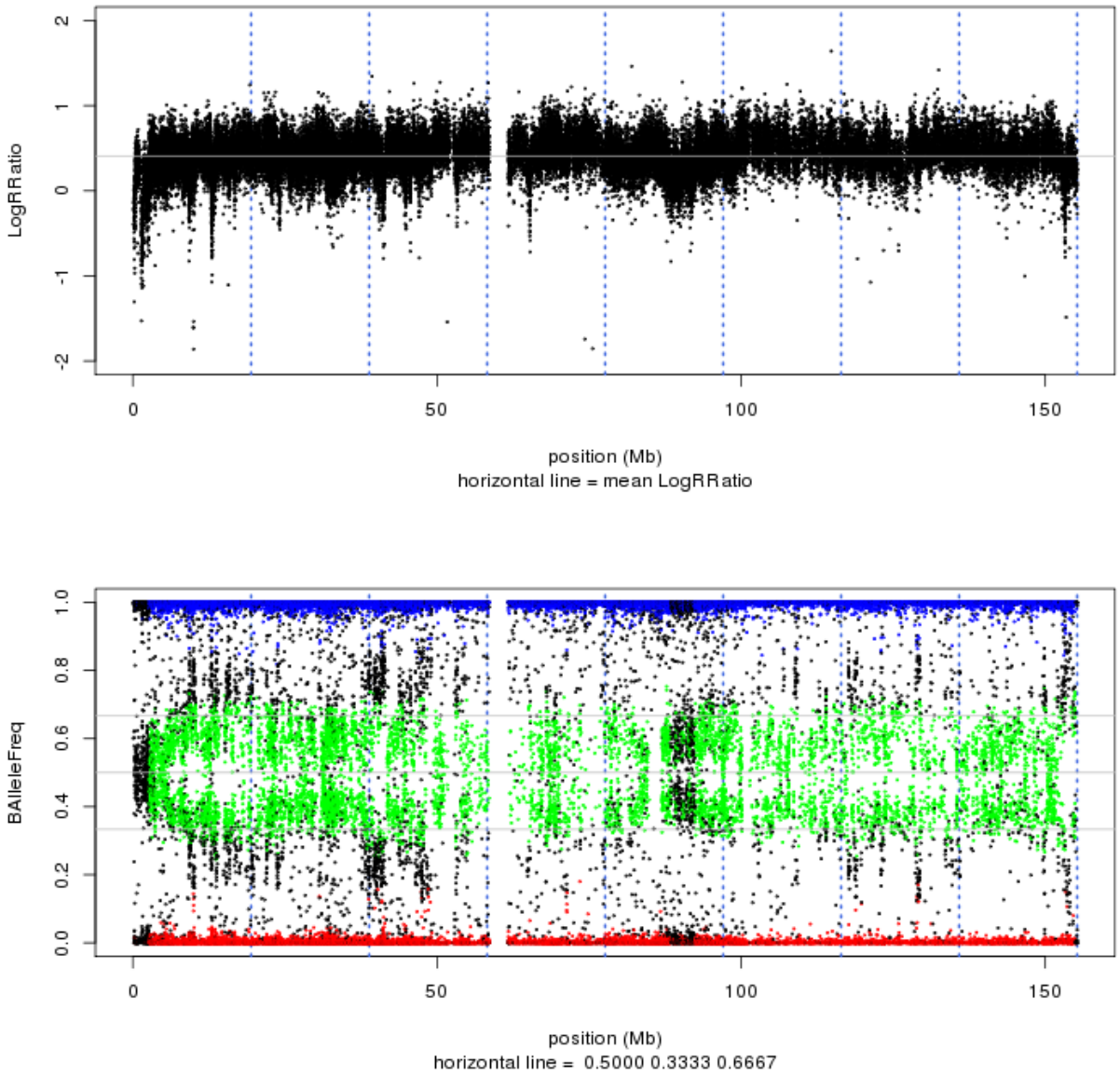
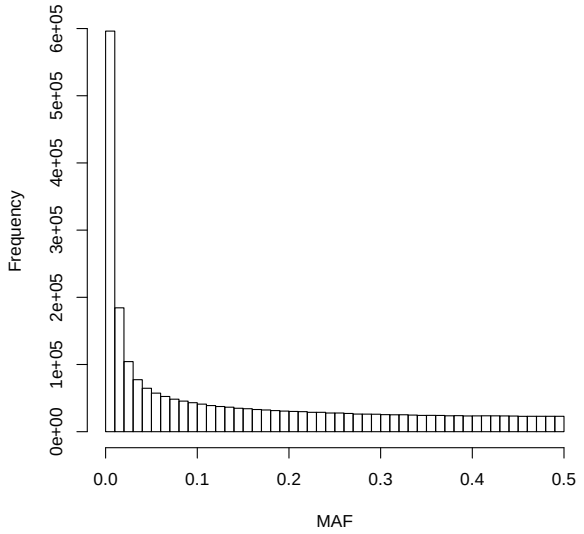
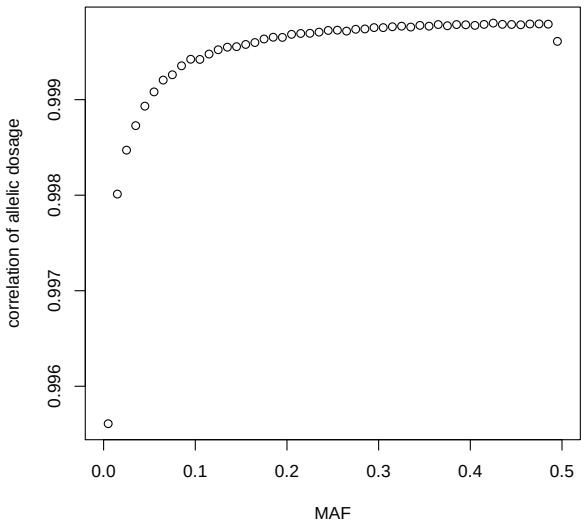


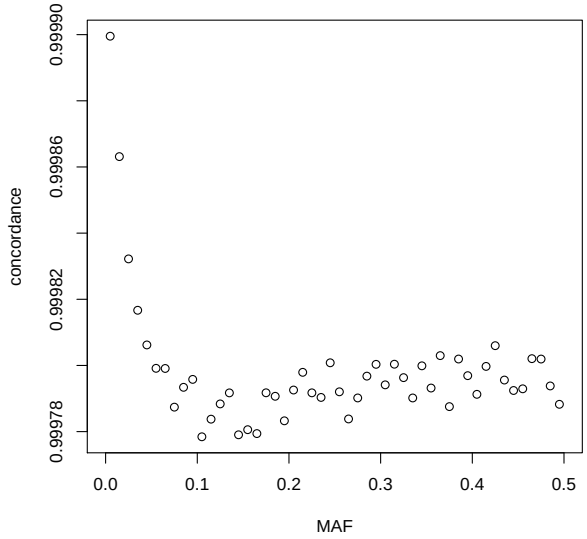
Figure 22: LRR and BAF plots for chromosome X in Female F. This chromosome shows a genotyping artifact on the X chromosome sometimes seen in Buccal samples. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing).



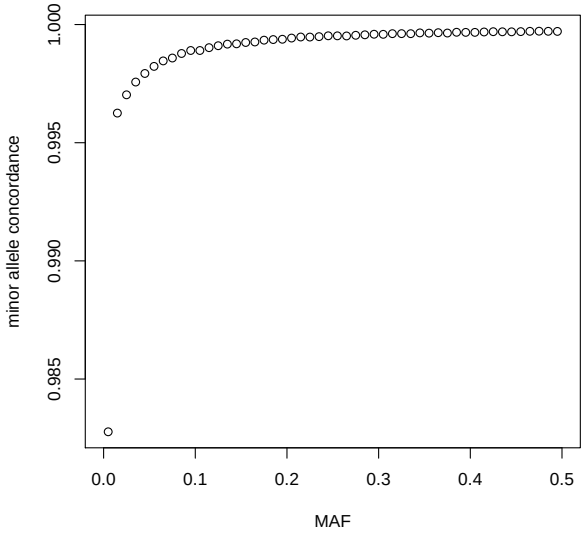
(a) Distribution of minor allele frequency.



(b) Correlation of allelic dosage.



(c) Overall concordance.



(d) Minor allele concordance.

Figure 23: Summary of concordance by SNP over 423 duplicate sample pairs, binned by minor allele frequency.

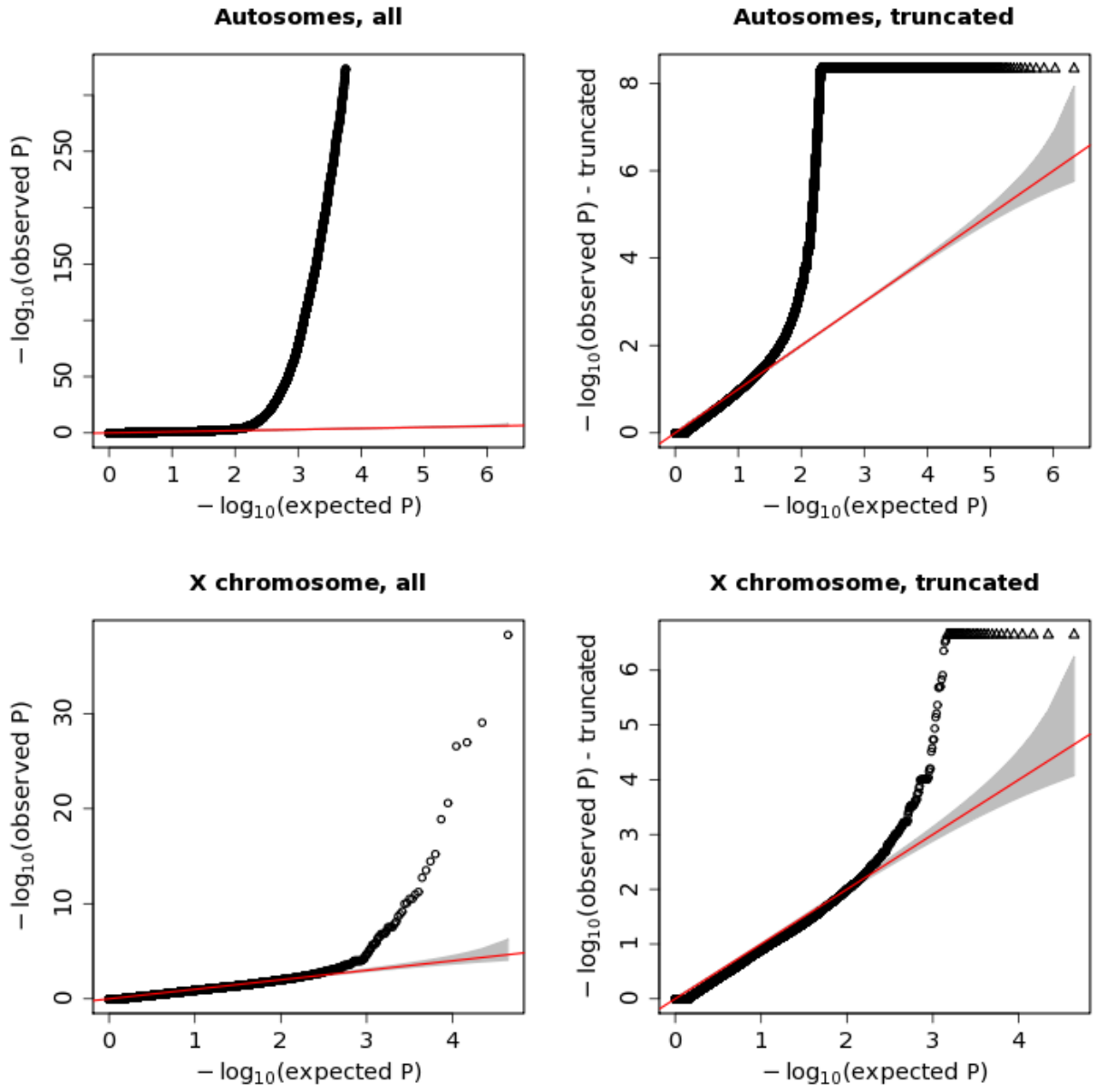


Figure 24: Quantile-quantile plots for $-log_{10}(p)$ from Fisher's exact test of Hardy-Weinberg equilibrium in European-ancestry subjects. Plots in the left column show all SNPs, whereas those in the right column have the Y-axis truncated to show more clearly the point of deviation from expectation.

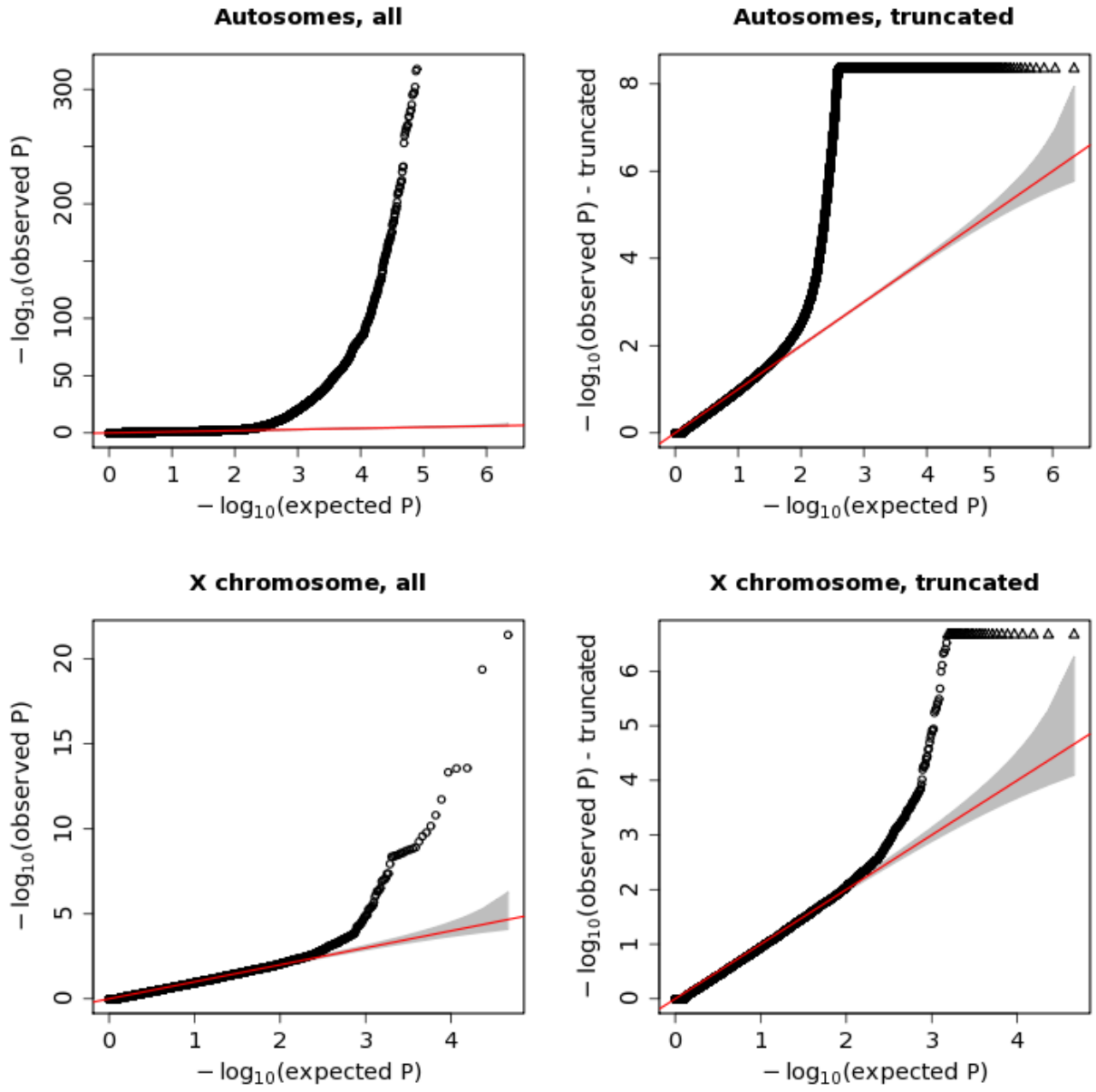
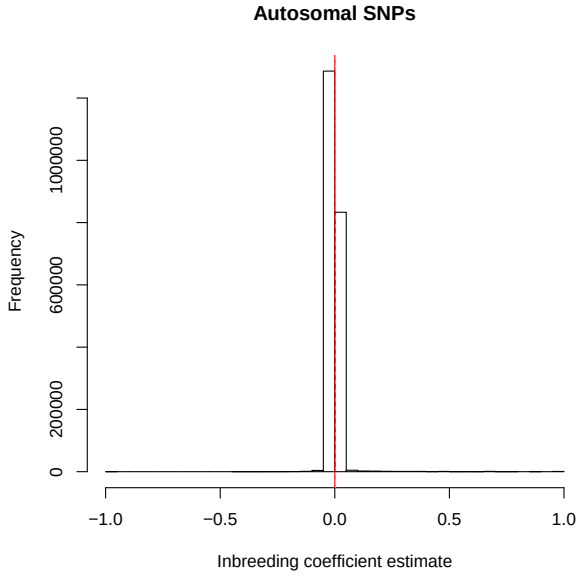
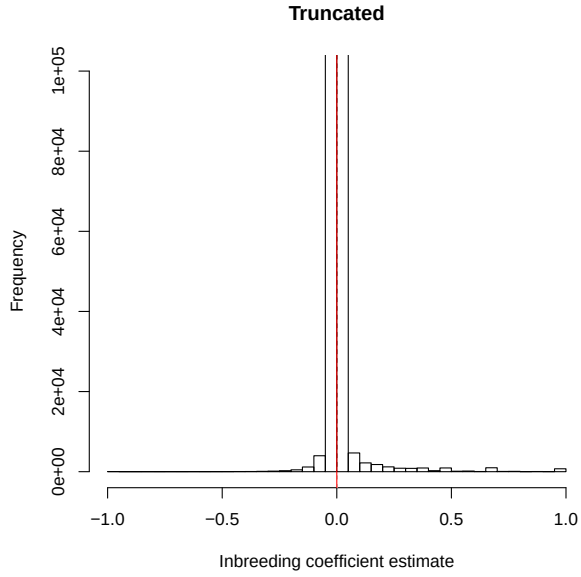


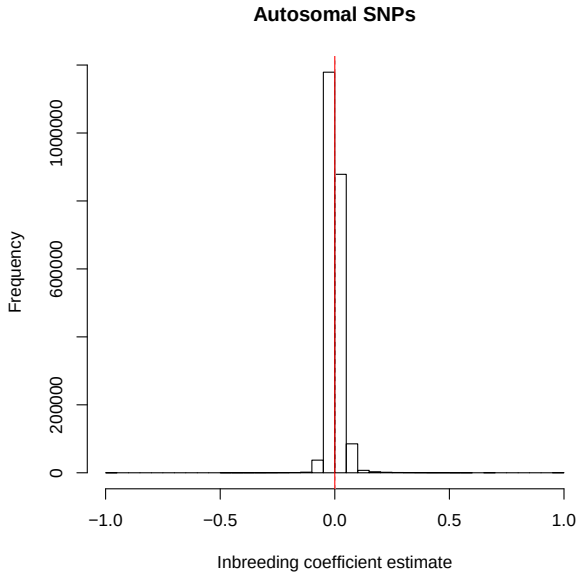
Figure 25: Quantile-quantile plots for $-log_{10}(p)$ from Fisher's exact test of Hardy-Weinberg equilibrium in African-ancestry subjects. Plots in the left column show all SNPs, whereas those in the right column have the Y-axis truncated to show more clearly the point of deviation from expectation.



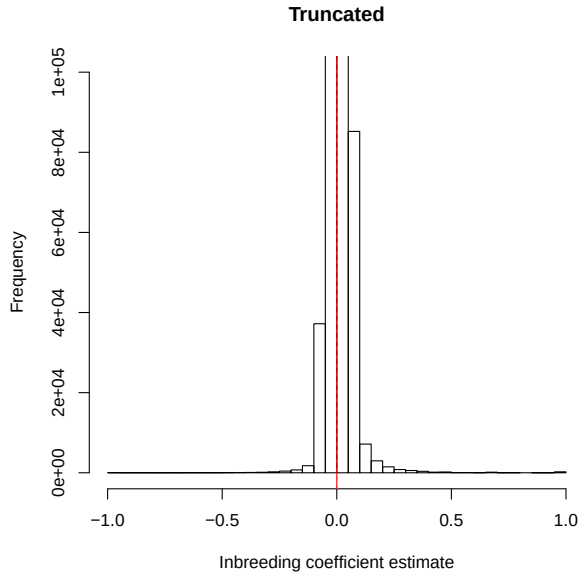
(a) European, all autosomal



(b) European, truncated

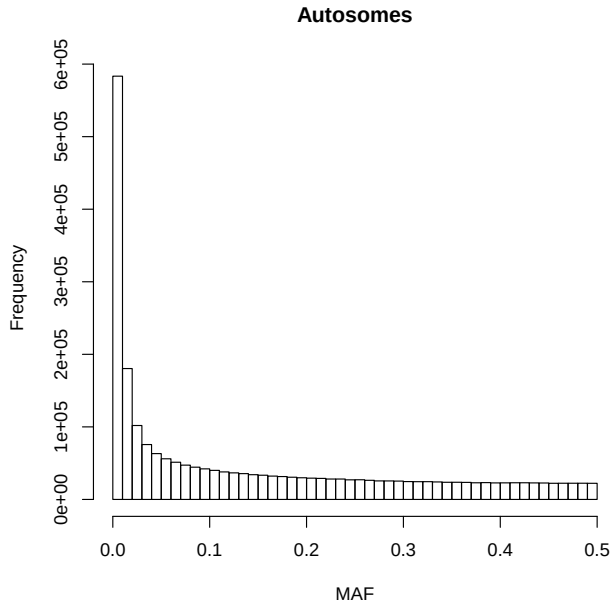


(c) African, all autosomal

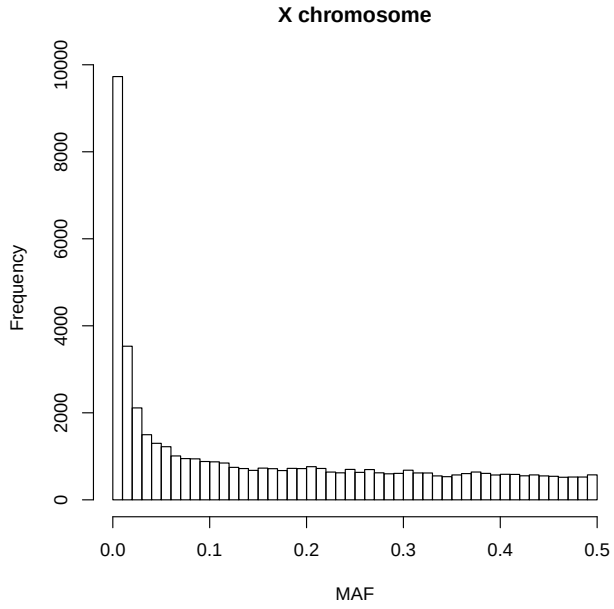


(d) African, truncated

Figure 26: Distribution of estimated inbreeding coefficient for all autosomal SNPs. The values range from -1 to 1.



(a) Autosomes



(b) X chromosome

Figure 27: Minor allele frequency distribution across all unrelated study subjects.