# Quality Control Report for Genotypic Data

University of Washington

September 4, 2013

**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 3,175 study subjects were genotyped on the Illumina HumanOmni2.5-8v1 array. The median call rate is 99.9% and the error rate estimated from 71 pairs of study sample duplicates is $1 \times 10^{-5}$. Genotypic data are provided for all subjects and SNPs. Generally, we recommend selective filtering of genotypic data prior to analysis to remove large ($> 5$ Mb) chromosomal anomalies showing evidence of genotyping error and to remove whole samples with an overall missing call rate $> 2\%$. In this study, there are four such anomalies and all samples have a 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 of work, health, social, psychological, family and economic status through 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. Similarly, new cohort households for 2010 were randomly assigned into one of these two groups with half being asked to complete an EFTF interview. Respondents who were not interviewed or did not consent to saliva collection in 2006 were asked to contribute a sample in 2010; over 50% consented to do so. This data release includes respondents who consented to the saliva collection in 2006 ('Phase 1'), 2008 ('Phase 2'), or 2010 ('Phase 3').

## 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. The HRS measures cognitive ability in several domains as well as depression. The study is supplemented with administrative linkages to Medicare claims files, providing diagnostic and utilization information, and to the National Death Index.

In addition to the core interview, HRS also conducts a number of supplemental studies, mainly in the form of mail and Internet surveys that are conducted between interview waves. These supplemental studies have been conducted since 1999 and have covered such topics as household spending, prescription drug use, diabetes treatment and self-management, disability vignettes, and parental investment in the human capital of their children.

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. The total recall score provided with these data is a masked version of the raw score (weighted averages across adjacent cells are used where cell size is less than 10).

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

DNA was extracted from saliva samples using Oragene. Thirty subjects who are duplicates from subjects genotyped in Phase 1 have DNA extracted from buccal swabs using the Qiagen Autopure method. A further thirty subjects who are duplicates from Phase 2 subjects have DNA extracted from saliva for a total of 60 cross-phase duplicate samples to be used for QC. The samples were genotyped in batches corresponding to 96-well plates. Each plate contained either one or two HapMap controls, as well as an average of 1.8 study sample duplicates.

The DNA samples were genotyped at the Center for Inherited Disease Research (CIDR) using the Illumina HumanOmni2.5-8v1 array and using the calling algorithm GenomeStudio version 2011.1, Genotyping Module version 1.9.4 and GenTrain version 1.0. The SNP annotation used by CIDR is "HumanOmni2.5-8v1_C," which uses genome build 37/hg 19.

# 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.).

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

A total of 3,313 samples (including duplicates) from study subjects were put into genotyping production, of which 3,265 were successfully genotyped and passed CIDR's QC process (Table 2). The subsequent QA process identified three subjects who were unexpected duplicates with previously genotyped samples, and two subjects with unexpected relatedness. A further 14 subjects withdrew from the study during the QA process. The final set of scans to be posted include 3,246 study participants and 68 HapMap controls. The 3,246 study scans derive from 3,175 subjects and include 71 pairs of duplicate scans (Table 3). The 68 HapMap control scans derive from 36 subjects, of which 32 are replicated two or more times. The study subjects occur as 3,171 singletons and two families of two members each. The study families were discovered during the analysis of relatedness (Section 9). The HapMap controls include nine trios (six of CEU and three of YRI).

# 6 Combined Phases 1-2 and Phase 3 Dataset

Subjects from the Health Retirement Study Phases 1 and 2 that were genotyped previously were combined with the current set of genotyped subjects ('Phase 3'). A set of genotype data for 15,708 combined subjects was created, which includes 2,365,472 overlapping SNPs from the Illumina HumanOmni2.5-4v1 array (Phases 1-2 genotyping) and the Illumina HumanOmni2.5-8v1 array (Phase 3 genotyping).

Phases 1 and 2 subjects were genotyped on the same array as the Phase 3 subjects but a different version. There are 74,277 SNPs that are unique to the HumanOmni2.5-4v1 array, and 10,953 SNPs unique to the HumanOmni2.5-8v1 array. Of the 2,368,902 overlapping SNPs, there are 3,430 SNPs with a different chromosome or position annotation. All 3,430 of the SNPs are those who have an unknown mapping or are mapped to the 'XY' chromosome in the HumanOmni2.5-8v1 array, and that are mapped to the X, Y or 'XY' chromosome in the HumanOmni2.5-4v1 array. All 3,430 of these SNPs are excluded from the combined posting, resulting in an overlap of 2,365,472 SNPs in common between the two array versions with the same rsID, chromosome and position annotations included in the "SNP_annotation.csv" and "SNP_analysis.csv" files. However, probes are synthesized in different batches for different versions of the array so that synthesis artifacts (such as probe failure) may differ even for a SNP with the same annotation in both array versions.

The median cross-study duplicate discordance is 0.6% among the 96 duplicate samples. Figure 1 shows the discordance rate for each pair, in increasing order. The high outlying sample is a pair of subjects whose missing call rates are 1.6% and 0.2%, for the genotyping done with Phases 1-2 and the current genotyping with Phase 3, respectively. Although this pair is an outlier, the discordance rate is still less than one percent.

Of the 15,708 combined Phases 1-2 and Phase 3 study subjects, a subset of 15,506 unrelated study subjects were used for principal components analysis (PCA). Figure 2 shows the results for the first four eigenvectors, colored by phase. Eigenvector three is highly correlated with phase, and almost perfectly separates the Phase 1-2 subjects and the Phase 3 subjects (colored in blue).

When rerunning PCA excluding 49,954 SNPs that had at least one discordant genotype call among the 96 duplicate samples, the division of the study samples by phase is no longer present. Figure 3 shows the batch effect is clearly eliminated when excluding the discordant SNPs. As a result, all further analysis with the combined dataset excludes the set of 49,954 discordant SNPs. A PLINK file for the combined datset includes 15,708 samples and 2,315,518 overlapping, non-discordant SNPs. The file "SNP_analysis.csv" contains a variable 'dup.scan.discord' which lists the number of discordant calls among the 96 duplicate samples. Those with an entry of 1 or larger are those that are discordant between the two batches and subsequently excluded from the combined PLINK file. On dbGaP, analysis results are provided for the combined dataset separately from the Phase 3 genotyped subjects.

# 7 Annotated vs. genetic sex

To check annotated vs. genetic sex, 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 4 shows two distinct clusters, with

one male falling within the female cluster. The outlier male sample has Y chromosome intensity characteristic of a male, but X chromosome intensity and X heterozygosity characteristic of a female. Further investigation of chromosomes X and Y in this subject (see Section 8) confirms this subject is a male XXY.

Furthermore, Figure 4 shows many females with higher than expected Y chromosome intensity; indeed, many females and males have similar Y chromosome intensities. When examining the subjects by DNA source, it becomes apparent that females with a higher Y chromosome intensity are those whose DNA from saliva was extracted using the Oragene method, as shown in Figure 5.

# 8 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. See [5] for a full description and application of this method. All sample-chromosome combinations with anomalies greater than 5 Mb or sample-chromosome combinations with the sum of the lengths of the anomalies greater than 10 Mb were verified by manual review of BAF and LRR plots.

Figure 6 shows BAF/LRR plots for chromosome 10 in Sample A. This chromosome shows a normal pattern, with LRR centered at 0 and BAF bands at 0, 0.5, and 1 (corresponding to AA, AB, and BB genotypes). Figure 7 shows BAF/LRR plots for chromosome 17 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. We interpret this anomaly as mosaic uniparental disomy due to mitotic recombination and we recommend filtering these genotypes.

Figure 8 shows BAF/LRR plots for chromosome 20 in Sample B, which shows a split in the heterozygous band in BAF with a decrease in LRR. This chromosome shows a pattern consistent with a mosaic deletion and we recommend filtering the SNPs in the anomaly for this sample.

Figure 9 shows the male sample identified as XXY during the sex check in Section 7. The X chromosome shows a disomic pattern for most SNPs, but a trisomic pattern for the pseudoautosomal (PAR) regions, indicating an XXY genotype.

There were a total of 11 acquired autosomal anomalies >5Mb in length, four of which are recommended for filtering. Generally, we recommend filtering out genotypes within all anomalies of this length that show indications of genotyping errors (i.e., a very wide split in the intermediate BAF band).

The breakpoints of all detected anomalies > 5 Mb in length are provided in the file "chromosome_anomalies.csv." Two PLINK files are provided: one with no filtering ("HRS_phase3_TOP_subject_level"), and one with genotypes in filtered anomaly regions set to missing ("HRS_phase3_TOP_subject_level_filtered").

We also examine BAF/LRR plots for evidence of sample contamination (more than 3 BAF bands on all chromosomes) and other artifacts. For this we examine scans that are high or low outliers for heterozygosity, high outliers for BAF standard deviation (for non-homozygous genotypes), and high outliers for relatedness connectivity (the number of samples to which a sample appears to be related with kinship coefficient $> 1/32$). No samples with evidence of contamination or unusual genotyping artifacts were found in this study.

# 9 Relatedness

The relatedness between each pair of participants was evaluated by estimation of the kinship coefficient ($KC$). The kinship coefficient ($KC$) for a pair of participants is

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

where $k2$ is the probability that two pairs of alleles are identical by descent (IBD) and $k1$ is the probability that one pair of alleles is IBD. Table 4 shows the expected coefficients for some common relationships. The $KC$ can be plotted in relation to the proportion of SNPs with zero identical by state (IBS), that is, the proportion of SNPs with opposite homozygous genotypes, to distinguish pairs of samples with differing levels of relatedness.

Subjects from the Phase 3 genotyping efforts were combined with previously genotyped Phases 1-2 in order to detect relatedness among Phase 3 subjects as well as relatedness across phases. As a result, 15,811 scans were used to detect relatedness, of which 3,216 are from the present Phase 3 genotyping.

IBD coefficients were estimated using 155,673 autosomal SNPs that were in common (and non-discordant) between the Phase 3 and Phases 1-2 genotyping, and the KING-robust procedure [6], but implemented in R using the package *SNPRelate*. The SNPs were selected using linkage disequilibrium (LD) pruning from all SNPs that are autosomal, non-monomorphic, non-discordant in the two genotyping sets, and had missing call rate $< 5\%$, with the constraint that no two SNPs are closer than 15 kb, and no two SNPs have an $r^2 > 0.1$. KING-robust was used because it is robust to population structure, which is needed for this mixture of European, African and Hispanic Americans. KING-robust provides estimates of the kinship coefficient and IBS0 (the fraction of SNPs that share no alleles), from which relationships can be inferred.

Figure 10 shows the detected relationships, color-coded by phase. The same plot is shown in Figure 11, color-coded by detected relationship type. All expected 30 Phase 1, 30 Phase 2 and 36 HapMap duplicates were found, for a total of 96 duplicate pairs. There were three Phase 3 subjects that unexpectedly were duplicates of subjects genotyped with Phases 1-2. All three samples were dropped from the Phase 3 genotyping set. Additionally, of the corresponding three subjects in the Phases 1-2 set, two were dropped from the Phases 1-2 set in the combined dataset, and the identity of the third was verified. Two parent-offspring pairs were found in which each involved one sample from Phase 3 and one sample from Phases 1-2. One member of each of these pairs was dropped from the Phase 3 genotyping. Within the Phase 3 genotyped subjects, two pairs of full siblings were identified. The relatedness of these subjects was confirmed from the Study Investigator and as a result, two families were created in the Phase 3 subjects.

A second relatedness analysis was performed on just the Phase 3 subjects. All relationships found within Phase 3 that were detected from the analysis combined with Phases 1-2 were confirmed in this analysis.

The IBD coefficient estimates for the families are provided in the file "Kinship_coefficient_table.csv." A full analysis of these data must take relatedness into account. *For an analysis that assumes all participants are unrelated, we recommend selecting one subject from each family unit, using "unrelated" in "Sample_analysis.csv."*

# 10 Population structure

To investigate population structure, we use principal components analysis (PCA), essentially as described by Patterson et al. [7], 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.

PCA was run using three sets of study samples. Two runs included study subjects from Phase 3 genotyping only, both with and without HapMap controls. The Phase 3 runs included 3,175 study and 1,213 HapMap subjects, and 3,173 subjects when including only unrelated Phase 3 study subjects. One set included all subjects from Phases 1-2 and Phase 3 genotyping, for a total of 15,506 unrelated study subjects.

We and others [8] 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 polymorphic inversion regions [8]. 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) $> 5\%$, and excluding any SNPs with a discordance between HapMap controls genotyped along with the study samples and those in the external HapMap data set. In addition, the 2q21 (LCT), HLA, 8p23, and 17q21.31 regions were excluded from the initial pool. Two sets of SNPs were used, where one set was calculated using all SNPs overlapping between the Phases 1-2 and Phase 3 genotyping, and another set using SNPs genotyped on the Phase 3 subjects alone. Using Phase 3 unrelated study subjects, the LD pruning process selected 178,020 SNPs with all pairs having $r^2 < 0.1$ in a sliding 10 Mb window. The LD pruning process, using Phases 1-2 and Phase 3 unrelated study subjects, selected 155,707 SNPs with all pairs having $r^2 < 0.1$ in a sliding 10 Mb window.

Figure 12 shows the PCA results using Phase 3 study subjects along with 1,213 HapMap3 subjects. Figure 12a displays a subset of study subjects only, and similarly Figure 12b shows the results for HapMap subjects only, for a less cluttered display of the results. As expected, the self-identified 'White, not Hispanic' study subjects generally cluster with the CEU and TSI HapMap subjects, although some (with Mexican American ethnicity) trail down towards the Asians. The self-identified 'Asian' subjects are clustered directly on the JPT and CHB HapMap subjects, and the self-identified 'Black' study subjects spread between the European HapMap and African HapMap subjects, tending towards the African HapMap subjects, in general. Subjects with Hispanic ancestry are mainly spread between the European and Asian HapMap controls.

The analysis of all 3,173 unrelated Phase 3 study subjects, without HapMaps, is shown in Figure 13. In this analysis, we found study subjects to generally cluster with other subjects of the same self-identified ethnicity and race. Figure 14 shows the percent variance accounted for by each eigenvector, which is relatively small for all eight eigenvectors shown. Eigenvector 1 accounts for about 5.5% of the variance, eigenvector 2 accounts for 1.5% of the variance, and the remaining eigenvectors account for less than 1%.

Figure 16 shows the results when analyzing Phases 1-2 and Phase 3 study subjects together. Similar to Figure 13, samples with the same self-identified ethnicity and race cluster together. Self-identified 'White' subjects span from the European cluster down towards samples with high Asian ancestry. Self-identified 'Black' subjects range from samples with high African ancestry towards those with mostly European ancestry. Subjects of other ancestries are scattered throughout.

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. Figures 17 and 18 show these plots for the first 8 eigenvectors for the Phase 3 and combined Phases 1-2 and 3, respectively. No clusters of highly correlated SNPs are evident in these plots, indicating that each eigenvector is related to many SNPs distributed across all chromosomes.

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 14 for Phase 3 and Figure 15 for combined PCA) and the association of each eigenvector with the recall score (Table 6 for Phase 3 and Table 7 for the combined PCA). In both the Phase 3 and combined PCA runs, the scree plot shows that the fraction of variance accounted for falls off dramatically after the first component, while the association tests indicate a significant relationship between recall score and the sixth eigenvector for Phase 3 ($p = 5e - 06$). The association tests indicate a significant relationship between recall score and the seventh eigenvector for the combined Phases 1-3 PCA results ($p = 7e - 06$). Thus, for the covariates used in the preliminary association tests described in Section 20, we choose eigenvectors 1-6 when considering just Phase 3 subjects and eigenvectors 1-7 when using the combined set of Phases 1-2 and Phase 3.

## 11 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 the Phase 3 study, the two missing rates by sample are very similar, with median values of 0.0012 (*missing.e1*) and 0.00087 (*missing.e2*). Figure 19 shows the distribution of *missing.e1*. All samples have a missing rate less than 2%.

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 0.0302% and 94.77% of SNPs have a missing call rate $< 1\%$.

*We recommend filtering out samples with a missing call rate* $> 2\%$ *(although there are none in this study) and SNPs with a missing call rate* $> 2\%$*.*

A missing call rate association with the trait of interest can lead to spurious genetic associations, since missingness is often nonrandom [9]. We tested for a such an association using linear regression of $log_{10}(missing.e1)$ on total recall score. The difference is not significant ($p = 0.227$).

## 12 Batch effects

The Phase 3 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 ($p < 2e-16$), but all plates have a low mean missing call rate (Figure 20).

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 21 shows the mean odds ratio for all plates. We concluded that there are no problematic plate effects.

## 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 $\alpha$ and $\beta$ 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 $\alpha$ and $\beta$ 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 $\alpha$ and three $\beta$ parameters), but here we assume they are the same. In this case, since the median discordance rate over all sample pairs is $2.0 \times 10^{-5}$, a rough estimate of the mean error rate is $1.0 \times 10^{-5}$ 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 8 shows these probabilities for x between zero and seven and $n = 103$. Here we chose $n = 103$ to correspond to the number of pairs of duplicate samples for both Phase 3 study and HapMap control samples. *We recommend a filter threshold of* $> 2$ *discordant calls because this retains* $> 98\%$ *of SNPs with an error rate* $< 10^{-3}$*, while removing* $> 33\%$ *of SNPs with an error rate* $> 10^{-2}$. This threshold eliminates 682 SNPs.

Figure 22 summarizes the concordance by SNP, binned by MAF. Figure 22a shows the number of SNPs in each MAF bin. Figure 22b shows the correlation of allelic dosage ($r^2$), which is greater for SNPs with higher MAF. Figure 22c shows the overall concordance, which is very high for all SNPs. For SNPs with low MAF, we expect high concordance because these SNPs are most likely to be called as homozygous for the major allele and thus be concordant by chance. Figure 22d 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.

## 14 Mendelian errors

Mendelian errors were analyzed in the nine trios of HapMap control subjects. Only 0.19% of SNPs have any Mendelian errors and just 634 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 Phase 3 study subjects who are (1) unrelated, (2) have missing call rate $< 2\%$, (3) self-identifed non-Hispanic white and (4) fall within 1 SD of all self-identified non-Hispanic whites for eigenvectors 1 and 2 in the PCA of all unrelated study subjects. Figure 23 shows quantile-quantile (QQ) plots for this HWE test.

A second HWE test used Phase 3 study subjects who are (1) unrelated, (2) have missing call rate $< 2\%$, (3) self- identified black and (4) fall within 1 SD of all self-identified blacks for eigenvectors 1 and 2 in the PCA of all unrelated study subjects. Figure 24 shows quantile-quantile (QQ) plots for this HWE test.

Both autosomal and X chromosome SNPs deviate from expectation between 0.01 and 0.001, although the X is closer to 0.001 or smaller. 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 are 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 25 shows the distribution of the inbreeding coefficient estimates for all autosomal SNPs. The distributions are roughly symmetric with mean=0.0015 and median=-7.23e-04 (European), mean=0.0011 and median=-0.0041 (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* $p = 0.0001$ *in either HWE tests 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 26 shows the distribution of minor allele frequency (MAF) for all unrelated study subjects. The percentage of all SNPs with MAF $< 2\%$ is 16.5% for the autosomes and 0.3% for the X chromosome.

# 17 Duplicate SNP probes

The Illumina HumanOmni2.5-8v1 array, with version C annotation, has 5,683 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 study samples. The numbers of pairs of duplicate SNPs with various levels of discordance are given in Table 9, along with the probability of observing each level given an assumed error rate (estimated as $1 \times 10^{-5}$ 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 5.0e-4% of the SNPs with an error rate of 0.0001 (∼twice the median),* 45% *of SNPs with error rate 0.001 and essentially all of the SNPs with error rate of 0.01. (This removes 107 pairs of SNPs.) We also recommend filters for SNPs involved in two annotation problems. (1) There are seven 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) One 1000 Genomes probe has incorrect A/B allele coding relative to the probe sequences. Finally, we recommend filtering out one member of each pair with* $\leq 6$ *discordant calls: the one with lowest missing call rate, since these provide redundant information. (This removes one member of each of the remaining 5,568 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 3,313 samples, of which 3,265 passed CIDR's QC process (Table 2). The subsequent data cleaning QA process identified 5 samples with unresolved identity issues. Fourteen subjects withdrew from the study during the course of the data cleaning. Therefore, 3,246 study scans will be posted on dbGaP.

In the combined Phases 1-2 and Phase 3 dataset as described in Section 6, a total 12,595 scans that passed the Phases 1-2 genotyping data cleaning process were combined with a total 3,211 scans genotyped in Phase 3 (Table 3). Two unexpected duplicates from Phases 1-2 were dropped as described in Section 9. Of the 30 Phase 1, 30 Phase 2, and 36 HapMap duplicate scans, one scan was chosen from each of the pairs. This results in a total of 15,708 subject scans in the combined dataset to be posted on dbGaP.

*In general, 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, which are provided in "Sample_analysis.csv." 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 in the Phase 3 study data. The suggested composite quality filter (for all rows except the final MAF$< 0.02$ 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 9.05% of the 2,379,855 SNP assays attempted.

In addition to the quality filter, we also suggest applying a minor allele frequency filter of 2% in unrelated study subjects in Table 1. The quality and MAF filters combined remove 31.51% of the SNP assays attempted. A majority (71.29%) of SNPs eliminated with the recommended quality filter have a MAF $< 2\%$.

A SNP filter for the combined Phases 1-2 and Phase 3 dataset was also generated. The union of the Phases 1-2 and the Phase 3 filters were used, excluding the MAF filter. MAF was calculated for the entire

set of combined samples together, and the monomorphic SNPs were filtered based on their MAF from the combined set. Table 10 summarizes the SNP filters in the combined dataset. The recommended combined filter removes 6.29% of the overlapping SNP assays.

*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

A linear regression model was used to obtain preliminary association test results for the Phase 3 data. The sample set consisted of 3,173 unrelated study subjects with *missing.e2*$<$ 0.02. Total recall score was regressed on each SNP's genotype score (coded as 0, 1, or 2), with age group, sex, and PCA eigenvectors 1-6 as covariates. In performing association tests for X-linked SNPs, male genotypes were coded as 0 and 2 (for BY and AY), whereas female genotypes were coded as 0, 1 and 2 (for BB, AB and AA). This coding seems appropriate to reflect the fact that, with X inactivation in females, the number of active alleles in homozygous females equals that in hemizygous males.

A second linear regression test obtained association test results using the combined Phases 1-2 and Phase 3 samples, resulting in 15,454 unrelated study subjects with *missing.e2*$<$ 0.02. The same model was used but also included eigenvector 7 as a covariate. Total recall score was regressed on the SNP genotype score, with age group, sex and PCA eigenvectors 1-7 as covariates. An additional covariate of Phase was also included, where this variable was 1 for subjects from Phases 1-2 and 2 for subjects from Phase 3.

Figures 27 and 28 show the QQ plots for the likelihood ratio test of the SNP effect, with no SNP filter, with the recommended quality filter, and with the quality plus expected heterozygosity filter in each of the regression tests. The corresponding Manhattan plots are shown in Figure 29 and Figure 30. No SNPs reach genome-wide significance in either association test.

Cluster plots for the top 9 hits from the model using Phase 3 subjects are shown in Figure 31 and the top 27 hits from this model are provided in the file "assoc_cluster_plots.pdf." Most of the SNPs in these plots show good clustering.

# 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, Tameka Shelford, Hua Ling, Elizabeth Pugh, and Marcia Adams

**Genetics Coordinating Center, Department of Biostatistics, University of Washington**
Caitlin McHugh, Stephanie M. Gogarten, Cathy Laurie, Bruce Weir and Quenna Wong

**dbGaP-NCBI, National Institutes of Health**
Nataliya Sharopova

## 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] Cathy C. Laurie, Cecelia A. Laurie, et al. Detectable clonal mosaicism from birth to old age and its relationship to cancer. *Nature Genetics*, 44:642–650, 2012.

[6] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

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

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

[9] 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.

Table 1: Summary of recommended SNP filters. The number of SNPs lost is given for sequential application of the filters in the order given. The CIDR technical filters are described in the text. The first two rows of the table are informativeness metrics, while the remaining rows are calculated quality metrics.

| Filter | SNPs lost | SNPs kept |
|---|---|---|
| SNP probes | | 2379855 |
| Intensity-only SNPs | 0 | 2379855 |
| CIDR technical filters | 30931 | 2348924 |
| MAF = 0 | 108739 | 2240185 |
| Duplicate SNPs | 5173 | 2235012 |
| Missing call rate >= 2% | 63011 | 2172001 |
| > 2 discordant calls in 103 study duplicates | 43 | 2171958 |
| > 1 Mendelian error | 473 | 2171485 |
| HWE P-value $< 10^{-4}$ | 6574 | 2164911 |
| Sex difference in allelic frequency $\geq 0.2$ | 331 | 2164580 |
| Sex difference in heterozygosity > 0.3 | 0 | 2164580 |
| Percent of SNPs lost due to quality filters | 9.05% | |
| MAF < 0.02 | 534626 | 1629954 |
| Percent of SNPs lost due to quality and MAF filters | 31.51% | |

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

| | Study | HapMap | Both |
|---|---|---|---|
| DNA samples into genotyping production | 3,313 | 68 | 3,381 |
| Failed samples | -48 | -0 | -48 |
| Scans released by genotyping center | 3,265 | 68 | 3,333 |
| Scans failing post-release QC | 0 | 0 | 0 |
| Scans with unresolved identity issues | -5 | 0 | -5 |
| Scans from subjects who withdrew from the study | -14 | 0 | -14 |
| Scans to post on dbGaP | 3,246 | 68 | 3,314 |

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

| | Study | HapMap | Both |
|---|---|---|---|
| Scans to post on dbGaP | 3,246 | 68 | 3,314 |
| Subjects | 3,175 | 36 | 3,211 |
| Replicated subjects | 71 | 32 | 103 |
| Families ($N > 1$) | 2 | 9 | 11 |
| Singletons | 3,171 | 9 | 3,180 |

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 | 2,314,174 | 256 | 7,436 | 51,968 | 3,657 | 2,364 |
| SNP tech failures | 0.0111 | 0.0312 | 0.1054 | 0.0661 | 0.0312 | 0.4179 |
| missing>0.05 | 0.0077 | 0.0161 | 0.0316 | 0.0005 | 0.0045 | 0.0029 |

Table 6: P-values for linear regression of recall score value on each of the first eight eigenvectors separately from PCA run using 3,173 unrelated Phase 3 study subjects.

| Eigenvector | P-value |
|---|---|
| 1 | <2e-16 |
| 2 | <2e-16 |
| 3 | 0.0030 |
| 4 | 0.0724 |
| 5 | 0.2325 |
| 6 | 5.6213e-06 |
| 7 | 0.0898 |
| 8 | 0.5419 |

Table 7: P-values for linear regression of recall score value on each of the first eight eigenvectors separately from PCA run using 15,506 unrelated Phases 1-3 study subjects.

| Eigenvector | P-value |
|---|---|
| 1 | <2e-16 |
| 2 | <2e-16 |
| 3 | 1.2126e-06 |
| 4 | 5.3453e-04 |
| 5 | 0.0397 |
| 6 | 0.8975 |
| 7 | 7.5956e-06 |
| 8 | 0.0203 |

Table 8: Probability of observing more than the given number of discordant calls in 103 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 > 2 discordant calls.

| | Assumed error rate | | | | |
|---|---|---|---|---|---|
| # discordant calls | 1.0e-05 | 1.0e-04 | 1.0e-3 | 1.0e-2 | # SNPs |
| >0 | 0.00206 | 0.0204 | 0.186 | 0.873 | 24277 |
| >1 | 2.1e-06 | 0.000207 | 0.0184 | 0.609 | 1870 |
| **>2** | **1.41e-09** | **1.39e-06** | **0.00122** | **0.335** | **682** |
| >3 | 7.06e-13 | 6.96e-09 | 6.02e-05 | 0.149 | 309 |
| >4 | 2.36e-16 | 2.75e-11 | 2.37e-06 | 0.0549 | 157 |
| >5 | 0 | 9e-14 | 7.72e-08 | 0.0171 | 82 |
| >6 | 0 | 2.77e-16 | 2.13e-09 | 0.00459 | 42 |
| >7 | 0 | 2.87e-17 | 5.11e-11 | 0.00108 | 19 |

Table 9: Probability of observing more than the given number of discordant calls in 5683 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 | 6.152603e-02 | 0.470073 | 0.998256 | 1.000000 | 1181 |
| >1 | 1.932188e-03 | 0.133527 | 0.987168 | 1.000000 | 361 |
| >2 | 4.065546e-05 | 0.026694 | 0.951930 | 1.000000 | 204 |
| >3 | 6.427449e-07 | 0.004093 | 0.877297 | 1.000000 | 146 |
| >4 | 8.135322e-09 | 0.000508 | 0.758782 | 1.000000 | 125 |
| >5 | 8.583362e-11 | 0.000053 | 0.608269 | 1.000000 | 115 |
| **>6** | **7.763000e-13** | **0.000005** | **0.449027** | **1.000000** | **107** |
| >7 | 6.157144e-15 | 0.000000 | 0.304665 | 1.000000 | 96 |
| >8 | 5.753572e-17 | 0.000000 | 0.190188 | 1.000000 | 93 |
| >9 | 1.460750e-17 | 0.000000 | 0.109520 | 1.000000 | 86 |
| >10 | 1.433567e-17 | 0.000000 | 0.058377 | 1.000000 | 81 |

Table 10: Summary of recommended SNP filters for the combined Phases 1-2 and Phase 3 dataset. The union of the SNP filters for Phases 1-2 and Phase 3 are taken, along with MAF as calculated on the combined dataset. SNPs with at least one discordant call of 96 duplicates are filtered.

| Filter | SNPs lost | SNPs kept |
|---|---|---|
| SNP probes | | 2365472 |
| Phases 1-2 SNP filter | 150893 | 2214579 |
| Phase 3 SNP filter | 52282 | 2162297 |
| >1 discordant call in 96 study duplicates | 37486 | 2124811 |
| MAF=0 | 49603 | 2075208 |
| Percent of SNPs lost due to quality filters | 6.29% | |
| MAF<0.02 | 638335 | 1436873 |
| Percent of SNPs lost due to quality and MAF filters | 35.12% | |

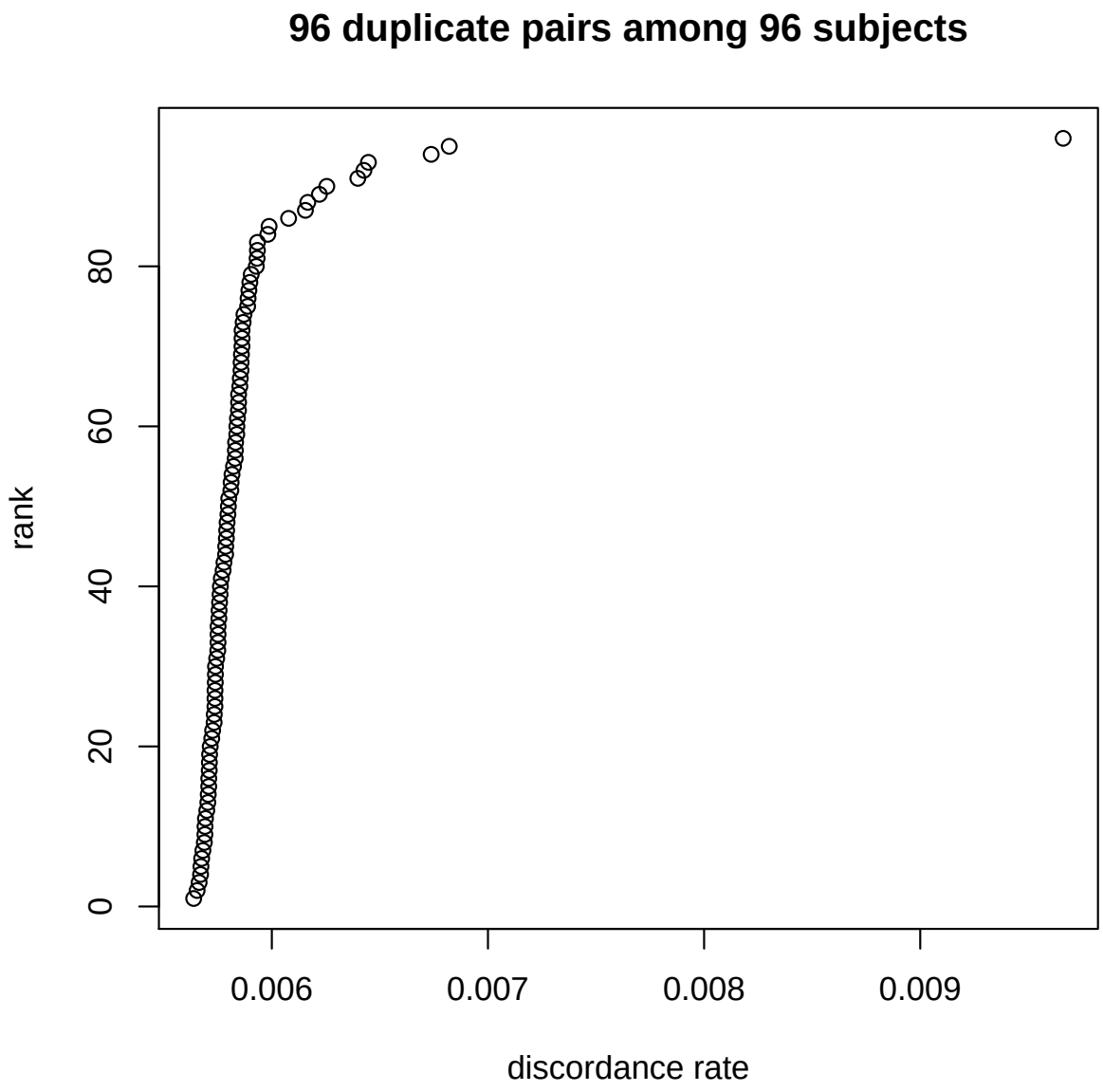



Figure 1: Duplicate sample discordance for 96 pairs of duplicates genotyped with Phases 1-2 and Phase 3. The median discordance is 0.6% among all 96 pairs.

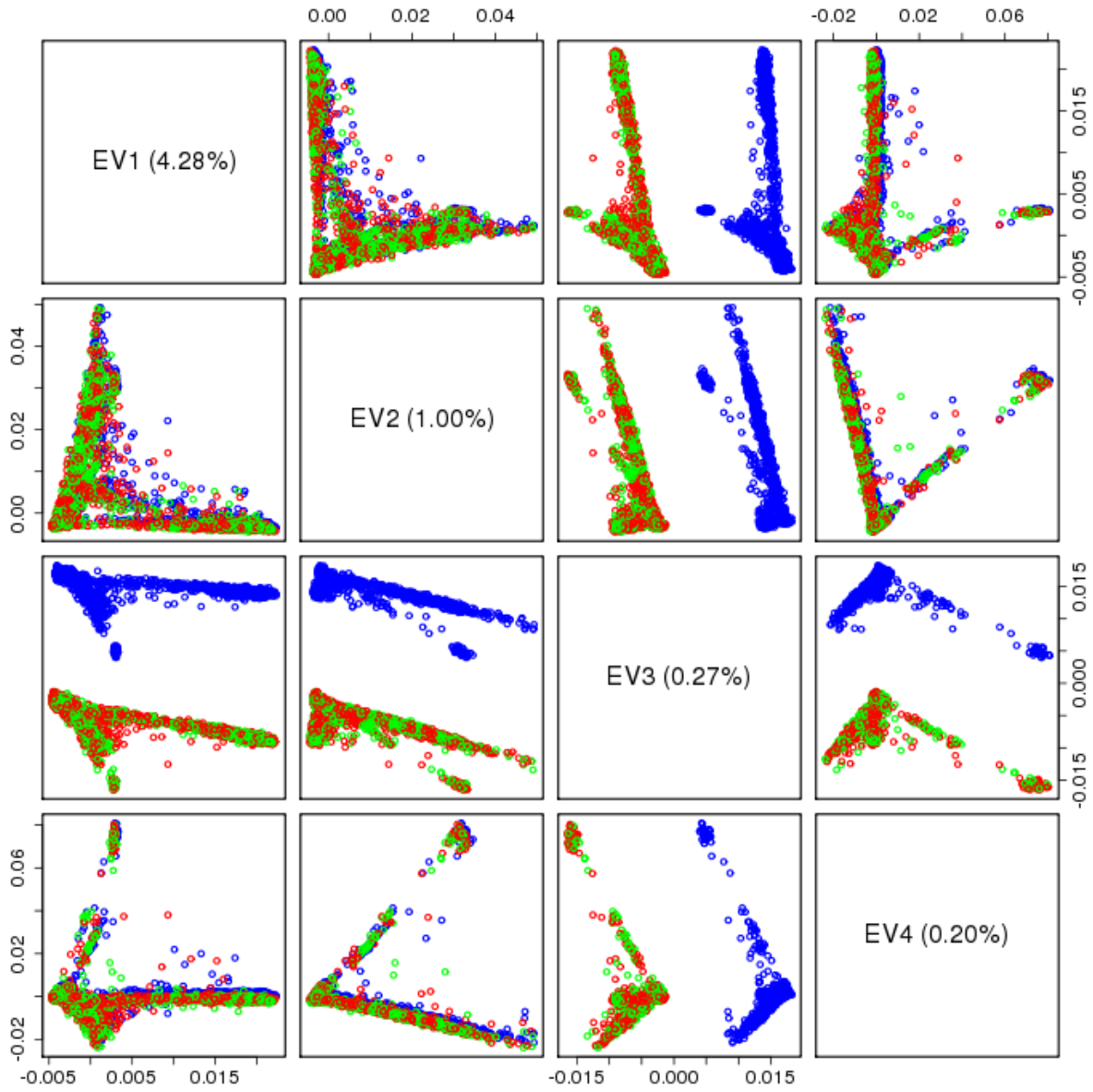



Figure 2: Principal component analysis of 15,627 Phase 1, 2 and 3 study subjects, colored by Phase. Phase 1 subjects are shown in red, Phase 2 subjects are green and Phase 3 subjects are colored blue. Eigenvector three is highly correlated with phase. Axis labels indicate the percent variation accounted for by each eigenvector.

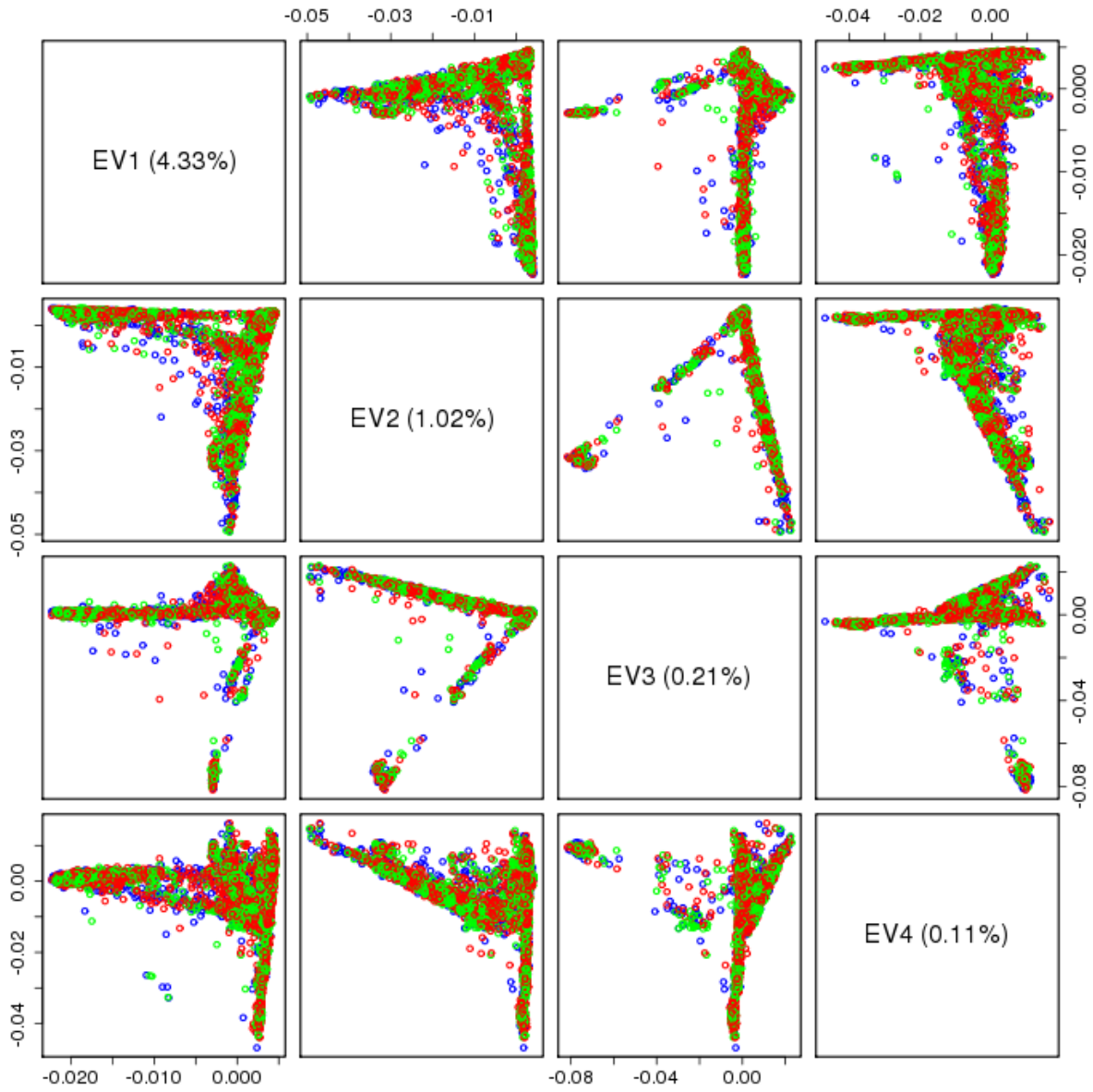



Figure 3: Principal component analysis of 15,627 Phase 1, 2 and 3 study subjects, colored by Phase, excluding 49,954 SNPs with at least one discordant call in 96 duplicate pairs. Phase 1 subjects are shown in red, Phase 2 subjects are green and Phase 3 subjects are colored blue. The batch effect is removed, as compared to Figure 2. Axis labels indicate the percent variation accounted for by each eigenvector.

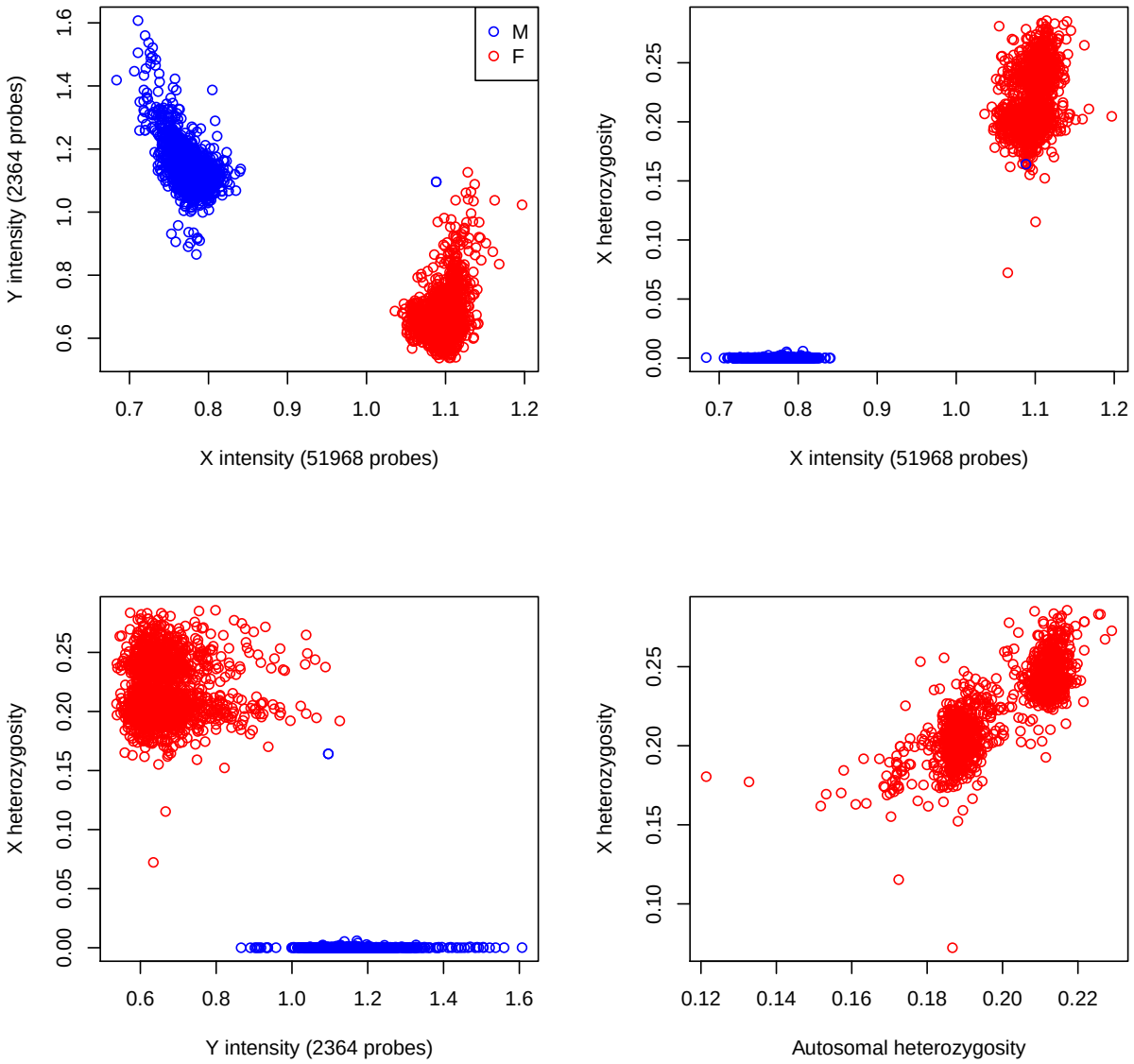



Figure 4: The X and Y intensities are calculated for each sample as the mean of the sum of the normalized intensities of the two alleles for each probe on those chromosomes. Sample sizes are given in the axis labels. X heterozygosity is the fraction of heterozygous calls out of all non-missing genotype calls on the X chromosome for each sample.

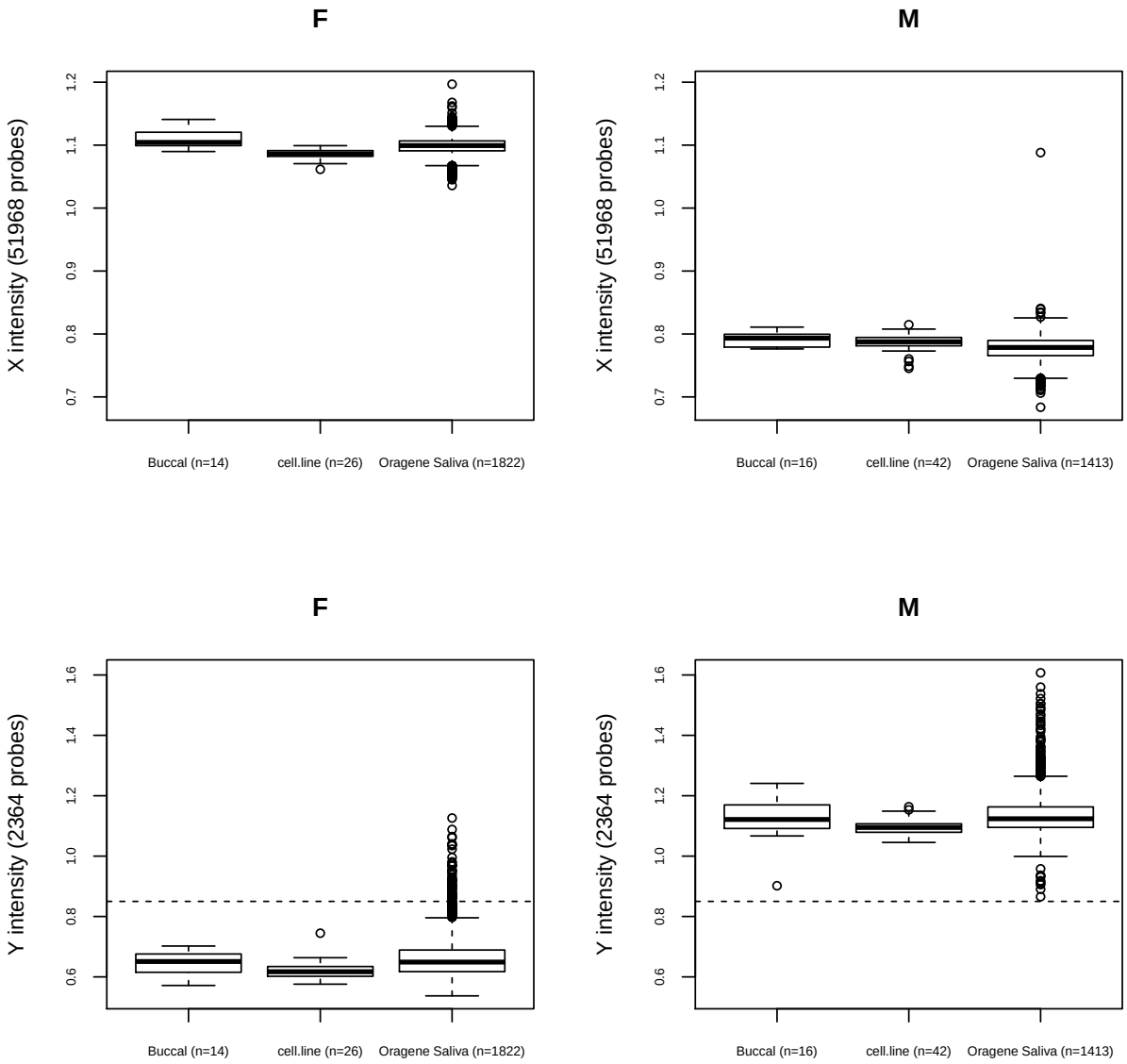



Figure 5: The X and Y intensities are calculated for each sample as the mean of the sum of the normalized intensities of the two alleles for each probe on those chromosomes. Sample sizes are given in the axis labels. X heterozygosity is the fraction of heterozygous calls out of all non-missing genotype calls on the X chromosome for each sample.

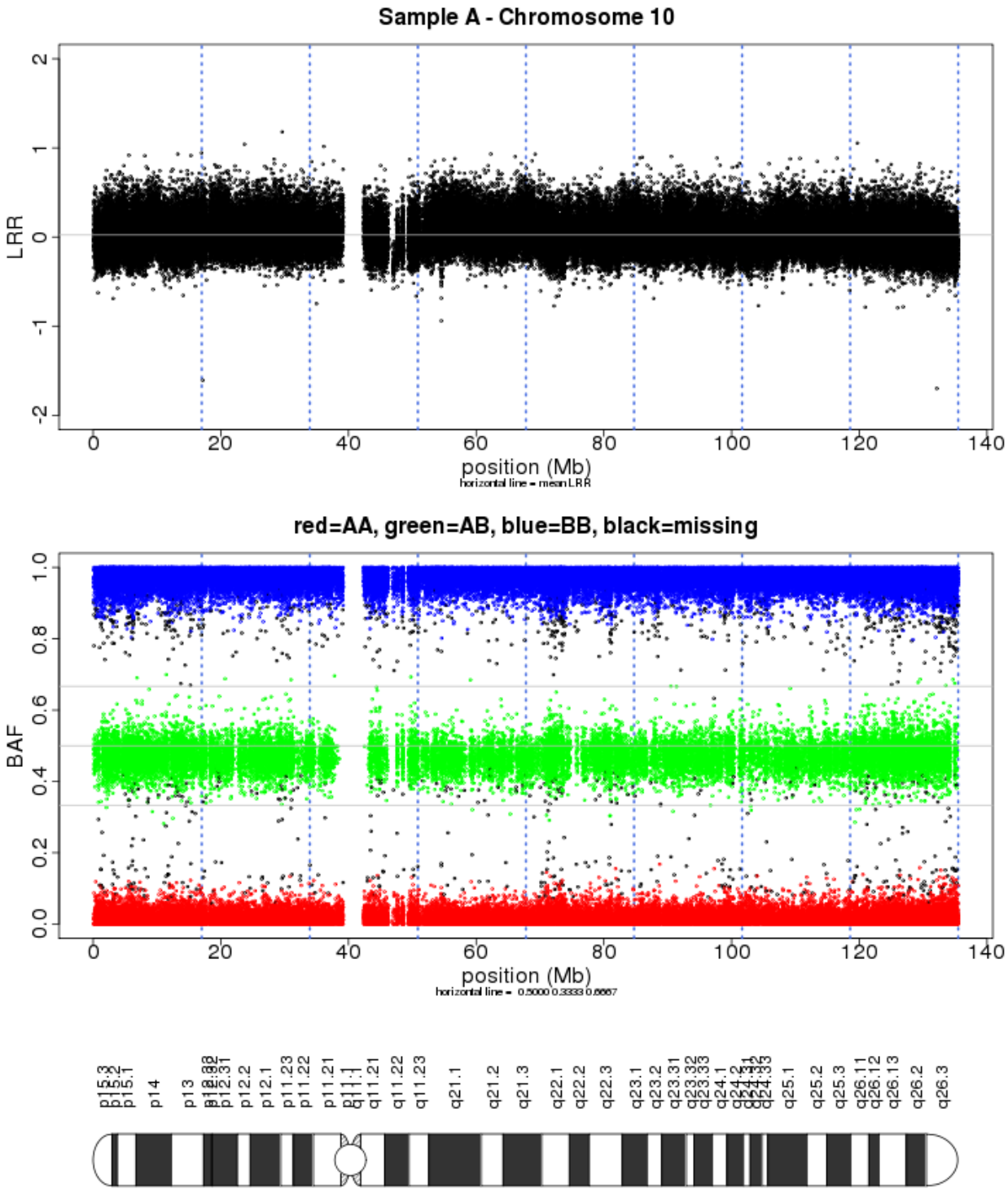



Figure 6: LRR and BAF plots for chromosome 10 in Sample A. This chromosome shows a normal pattern. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing).

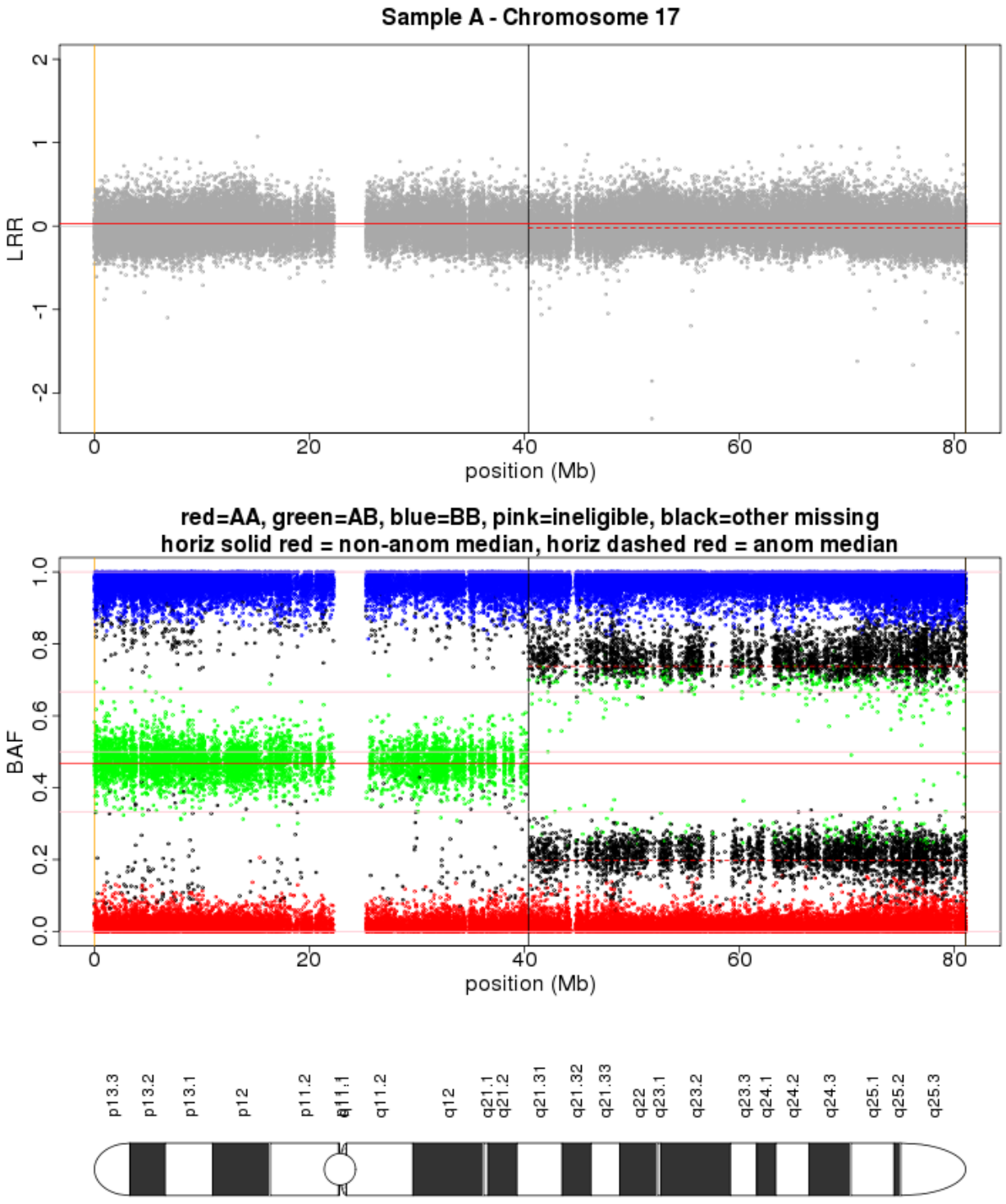



Figure 7: LRR and BAF plots for chromosome 17 in Sample A. This chromosome shows a terminal split in the heterozygous band wide enough to cause some heterozygous SNPs to be called as homozygous. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing). The horizontal solid red line 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.

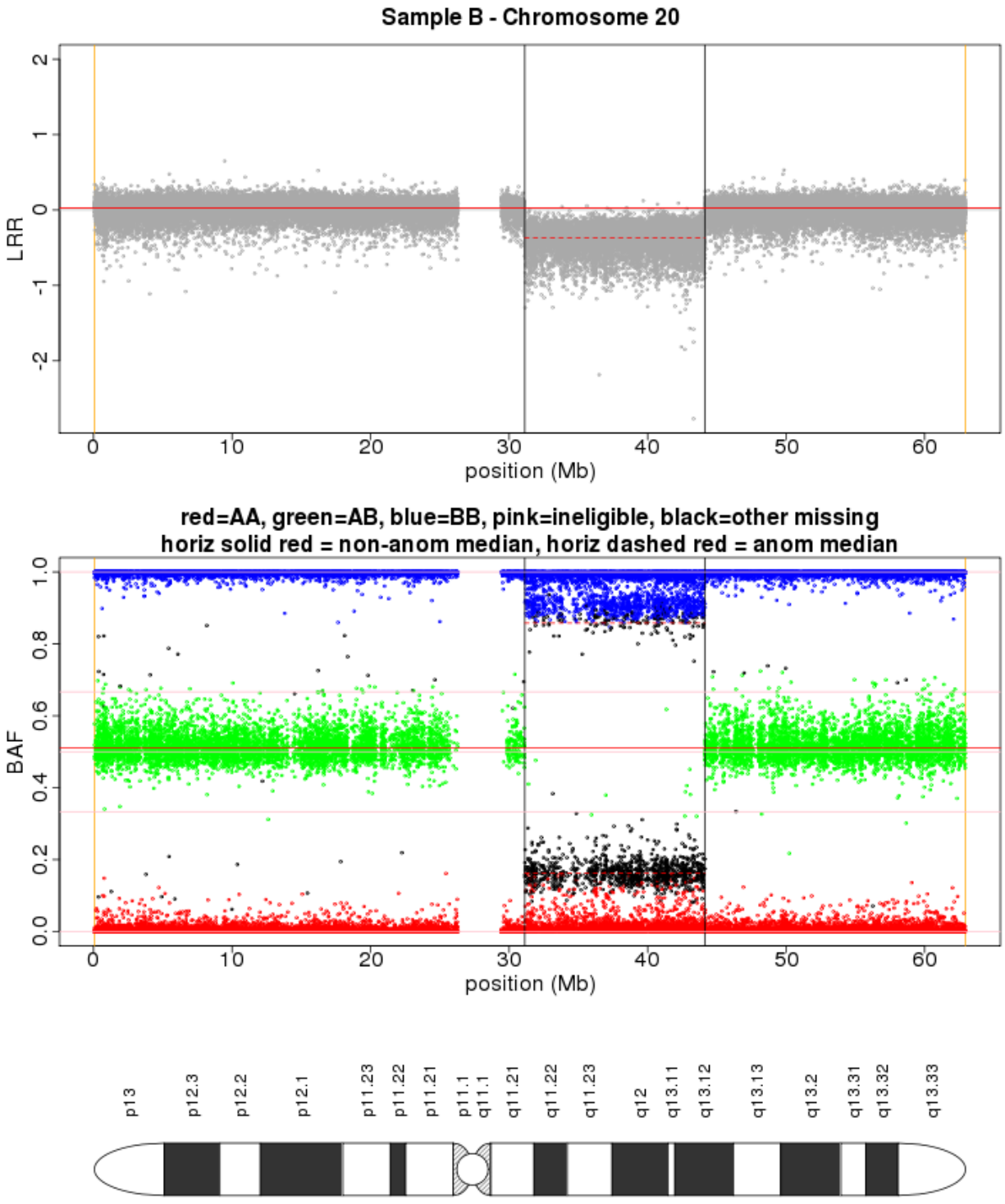



Figure 8: LRR and BAF plots for chromosome 20 in Sample B. This chromosome shows an interstitial split in the heterozygous band accompanied with a decrease in the LRR, indicating a deletion. Color-coding is for genotype calls (red=AA, green=AB, blue=BB, black=missing). The horizontal solid red line in 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.

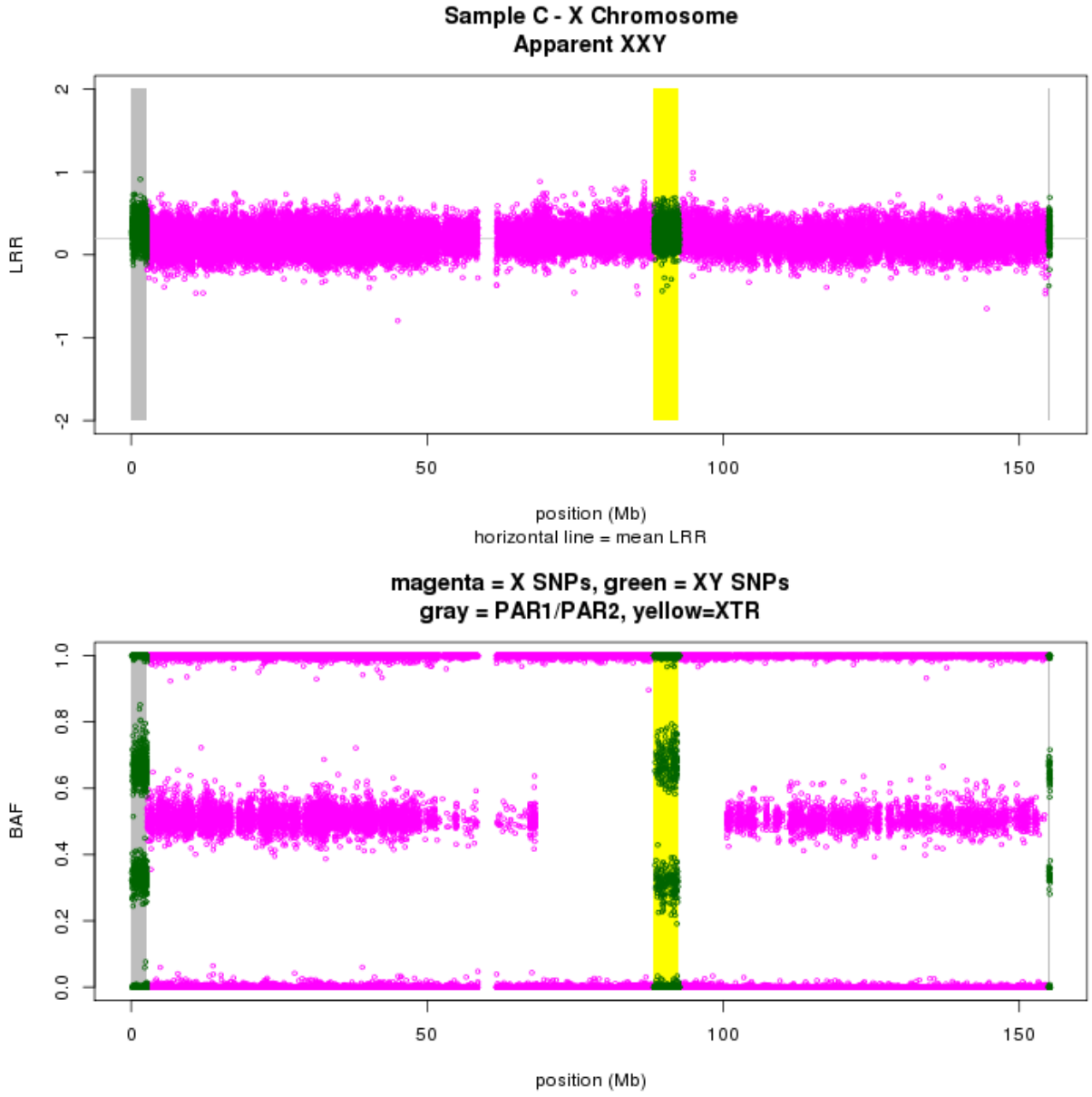



Figure 9: LRR and BAF plots for the X chromosome in Sample C. This chromosome shows a pattern consistent with an XXY karyotype. Color-coding is green for SNPs in the pseudo-autosomal regions (PAR1 and PAR2, shown in gray rectangles, and XTR, shown in a yellow rectangle) and pink for other X chromsome SNPs.

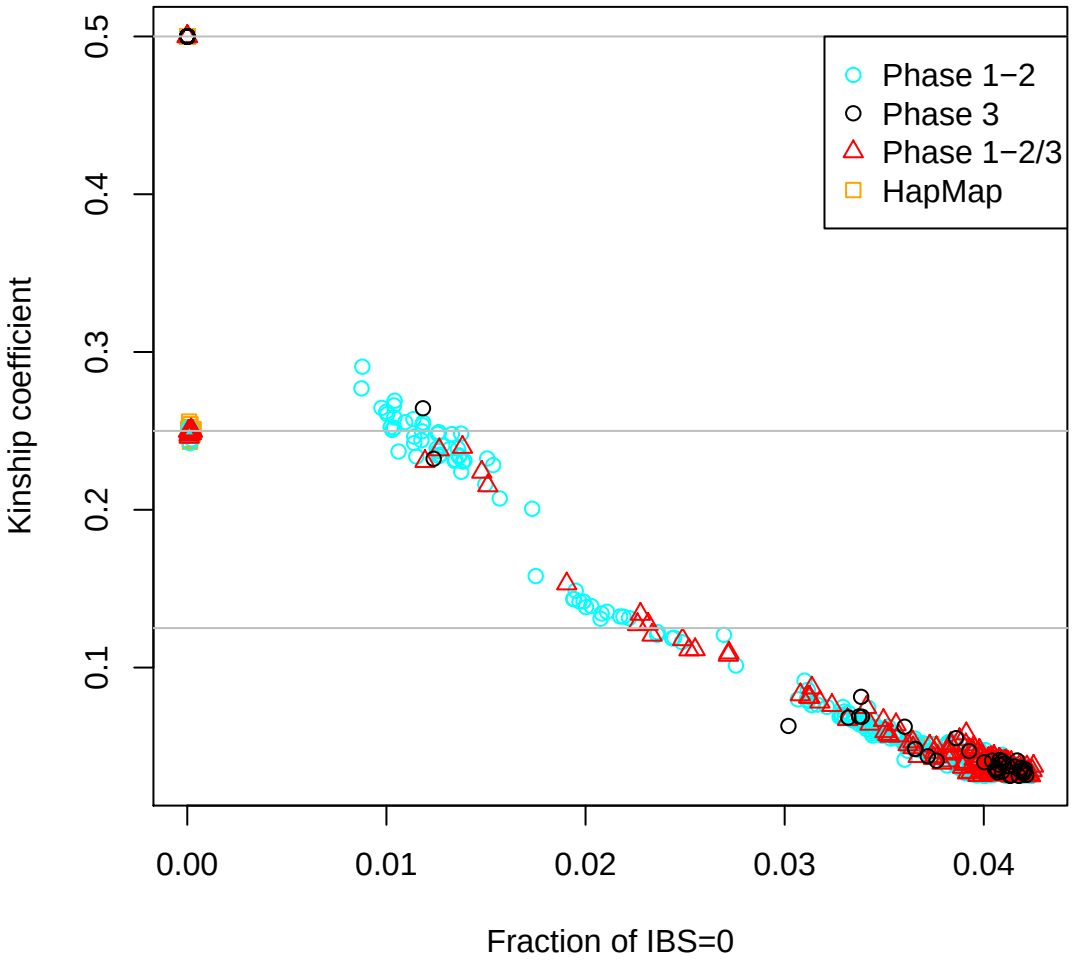



Figure 10: IBD coefficients to estimate relatedness. Each point represents a pair of samples. This plot shows 663 pairs of Phases 1-2 and Phase 3 study participants and HapMap controls with an estimated $KC > 1/32$, color-coded by whether the pair of samples are both from the same phase, are cross-phase or are HapMap samples. The X-axis shows the proportion of SNPs with zero IBS (e.g., AA and BB), while the Y-axis shows the kinship coefficient. Gray horizontal lines show the expected values for duplicates (KC=0.5), parent-offspring and full siblings (KC=0.25), and second-degree relatives (KC=0.125).

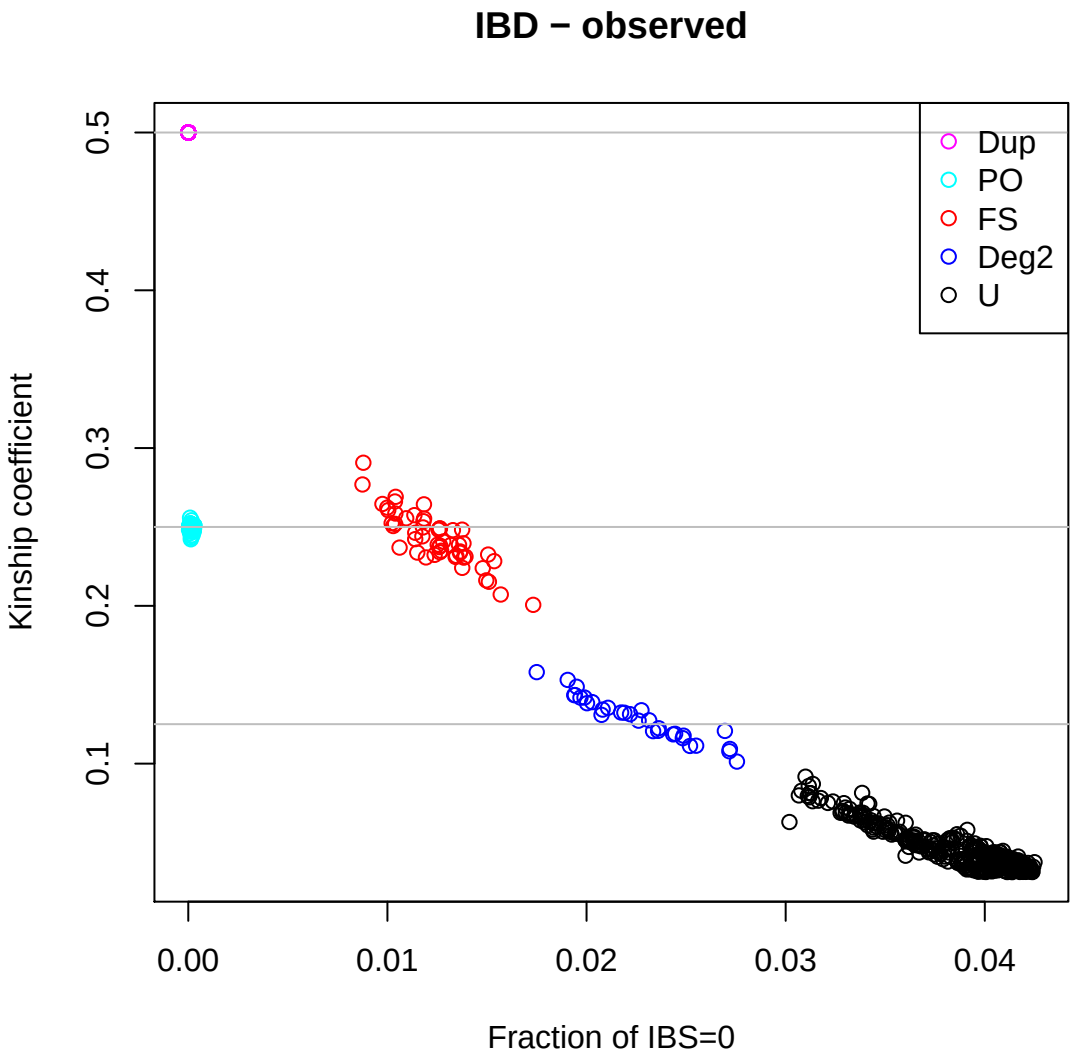
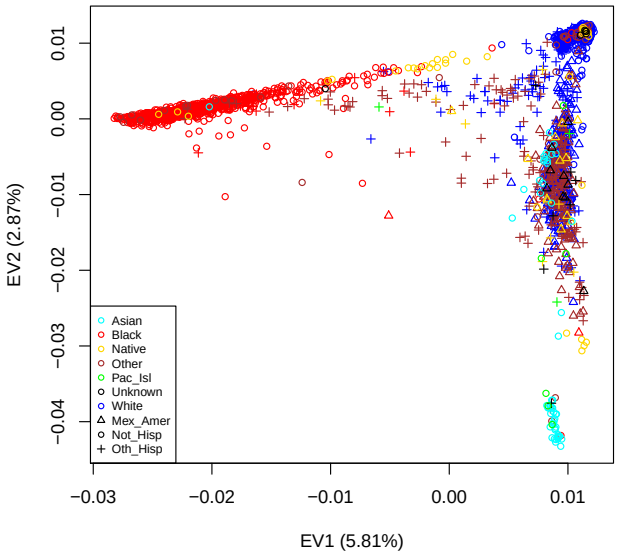
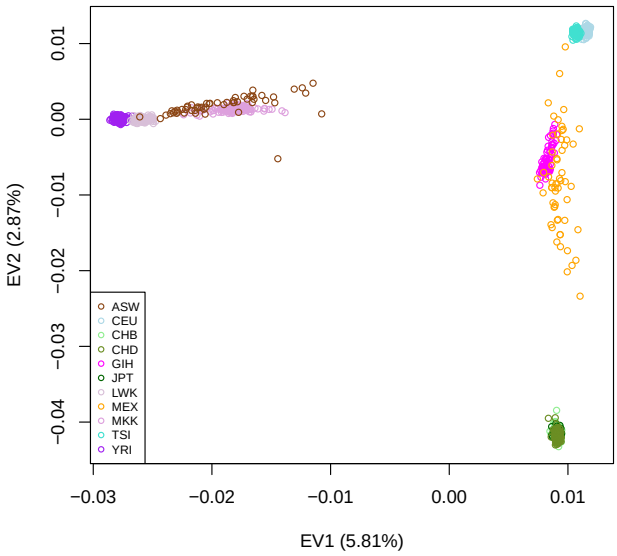



Figure 11: IBD coefficients to estimate relatedness. Each point represents a pair of samples. This plot shows 663 pairs of Phases 1-2 and Phase 3 study participants and HapMap controls with an estimated $KC > 1/32$, color-coded by detected relationship. In the legend, 'Dup' stands for duplicates, 'PO' means parent-offspring pairs, 'FS' annotates full-siblings, 'Deg2' denotes second degree relatives and 'U' stands for pairs of subjects related at third degree or higher. The X-axis shows the proportion of SNPs with zero IBS (e.g., AA and BB), while the Y-axis shows the kinship coefficient. Gray horizontal lines show the expected values for duplicates (KC=0.5), parent-offspring and full siblings (KC=0.25), and second-degree relatives (KC=0.125).




(a) Study Subjects




(b) HapMap Controls

Figure 12: Principal component analysis of 3,175 study subjects with HapMap controls. Figure 12a shows just study subjects, and Figure 12b shows just HapMap subjects. Color-coding is according to self-identified race, while symbol denotes ethnicity (Mexican American, Other Hispanic or not Hispanic). Axis labels indicate the percentage of variance explained by each eigenvector.

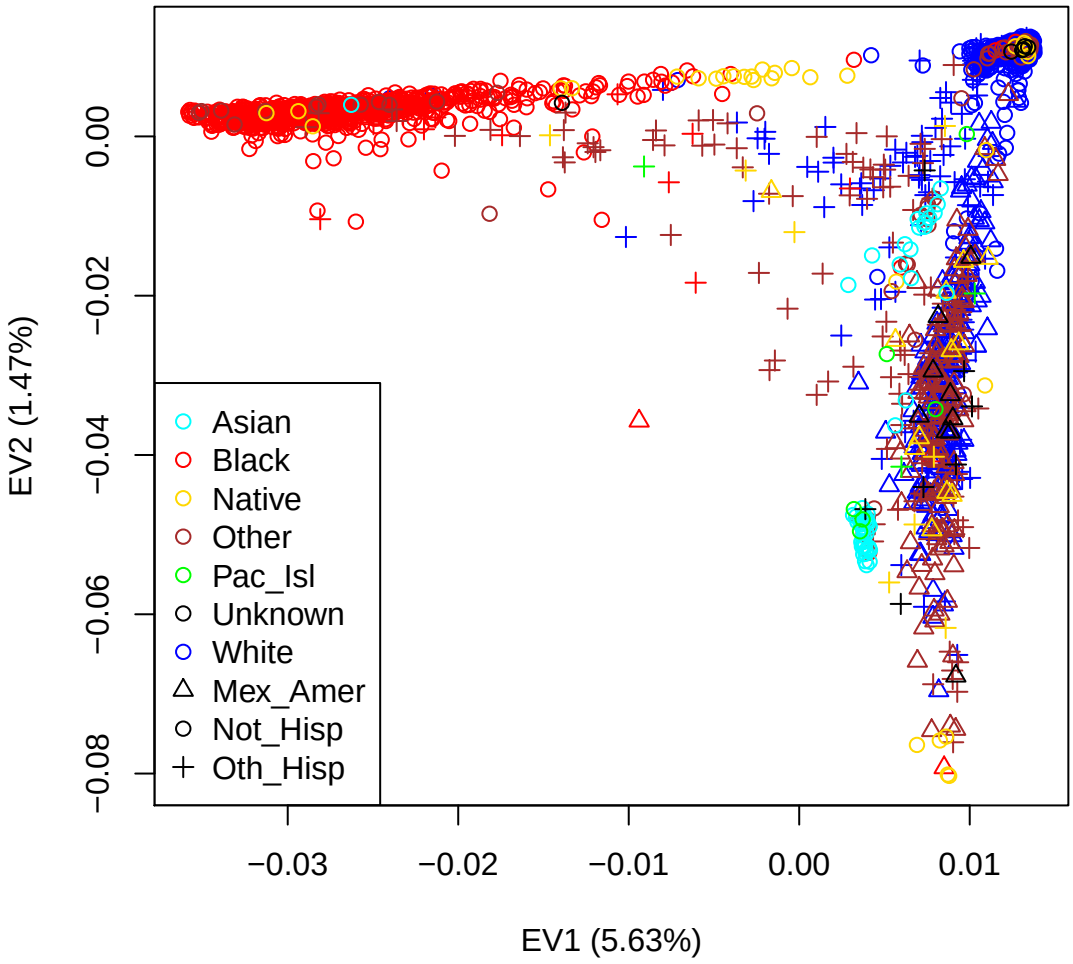



Figure 13: Principal component analysis of 3,173 unrelated study subjects without HapMap controls. Color-coding is according to self-identified race, while symbol denotes ethnicity (Mexican American, Other Hispanic or not Hispanic). Axis labels indicate the percentage of variance explained by each eigenvector.

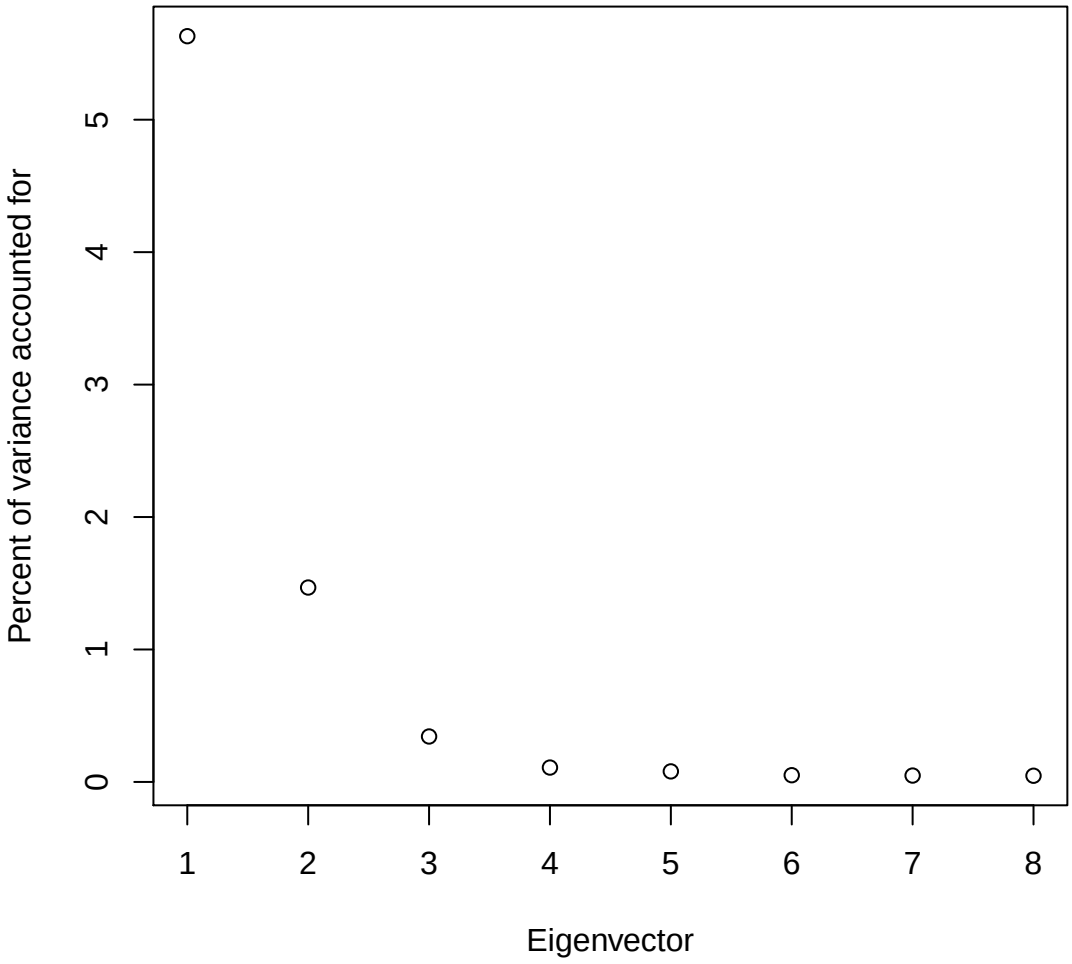



Figure 14: Scree plot for PCA shown in Figure 13.

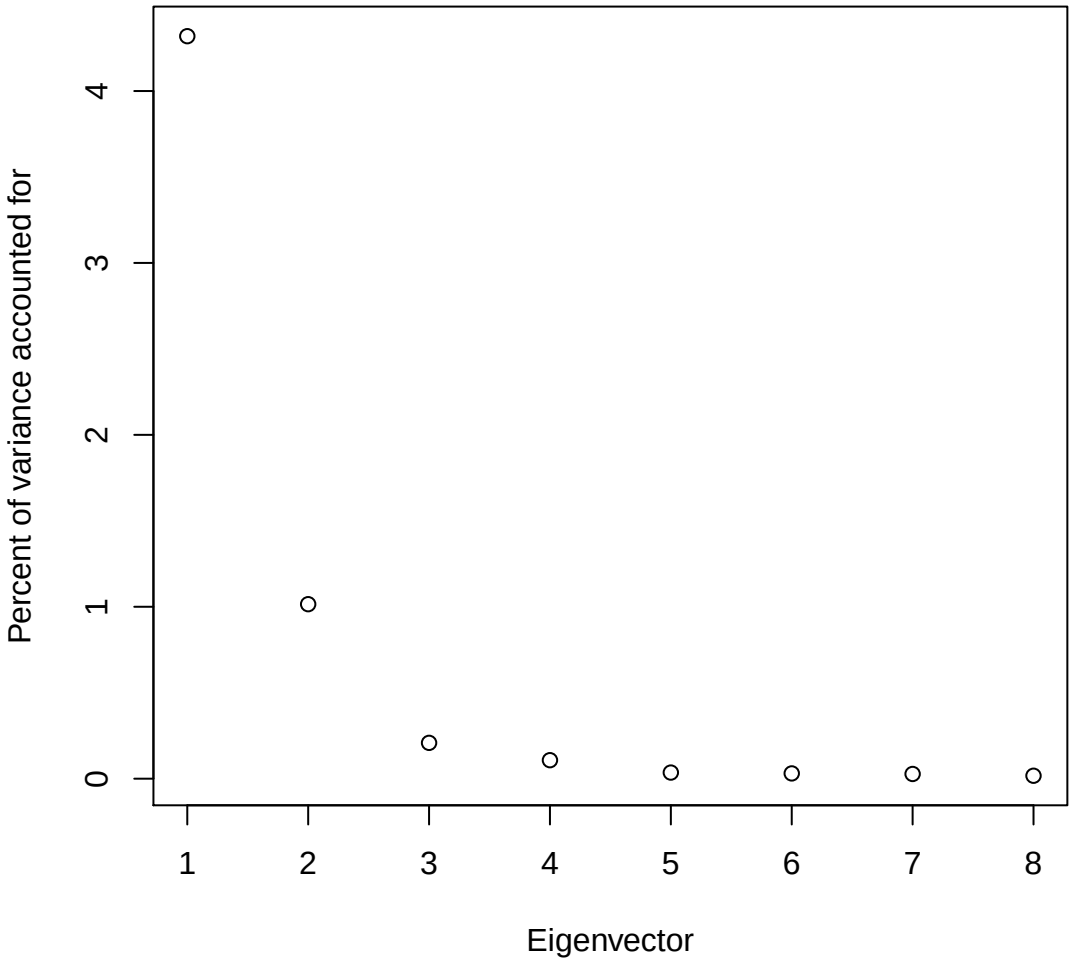



Figure 15: Scree plot for PCA shown in Figure 3.

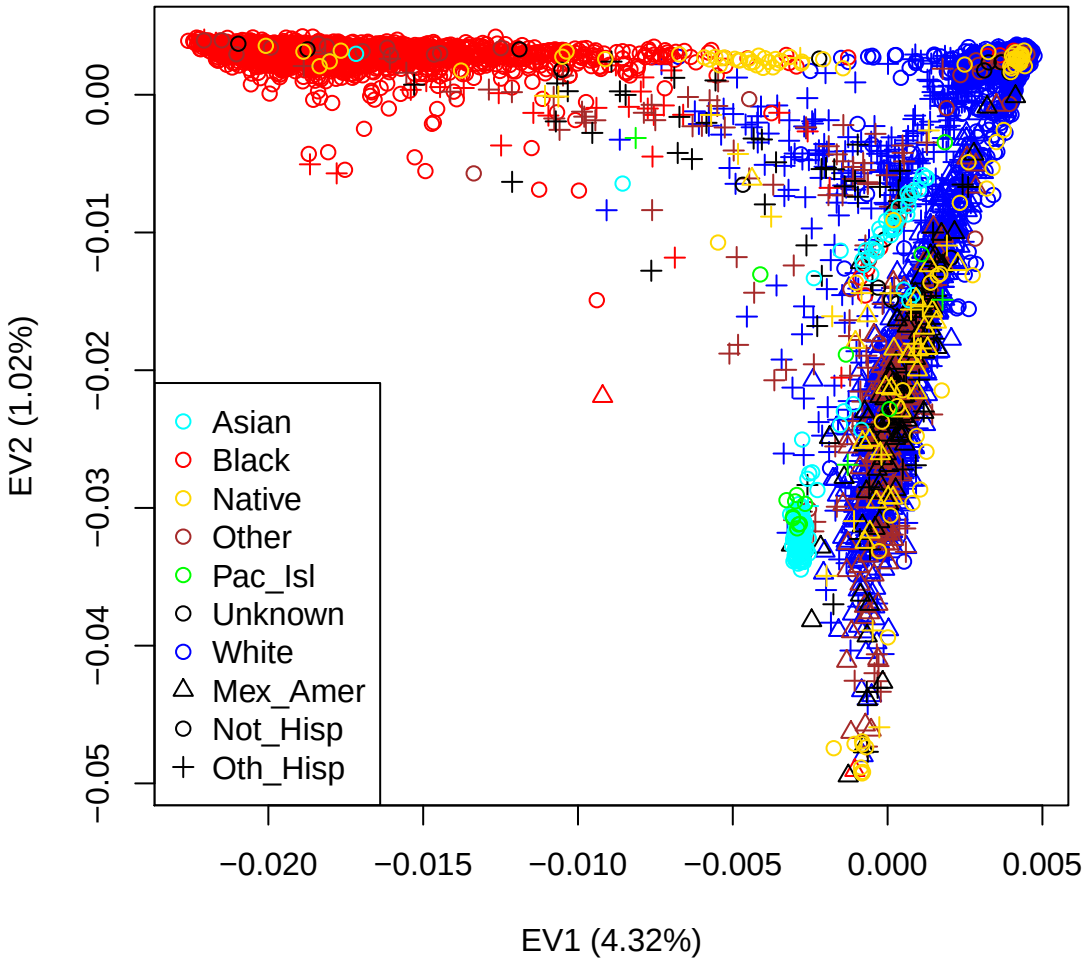



Figure 16: Principal component analysis of 15,506 unrelated study subjects from Phases 1-2 and Phase 3. Color-coding is according to self-identified race, while symbol denotes ethnicity (Mexican American, Other Hispanic or not Hispanic). Axis labels indicate the percentage of variance explained by each eigenvector.

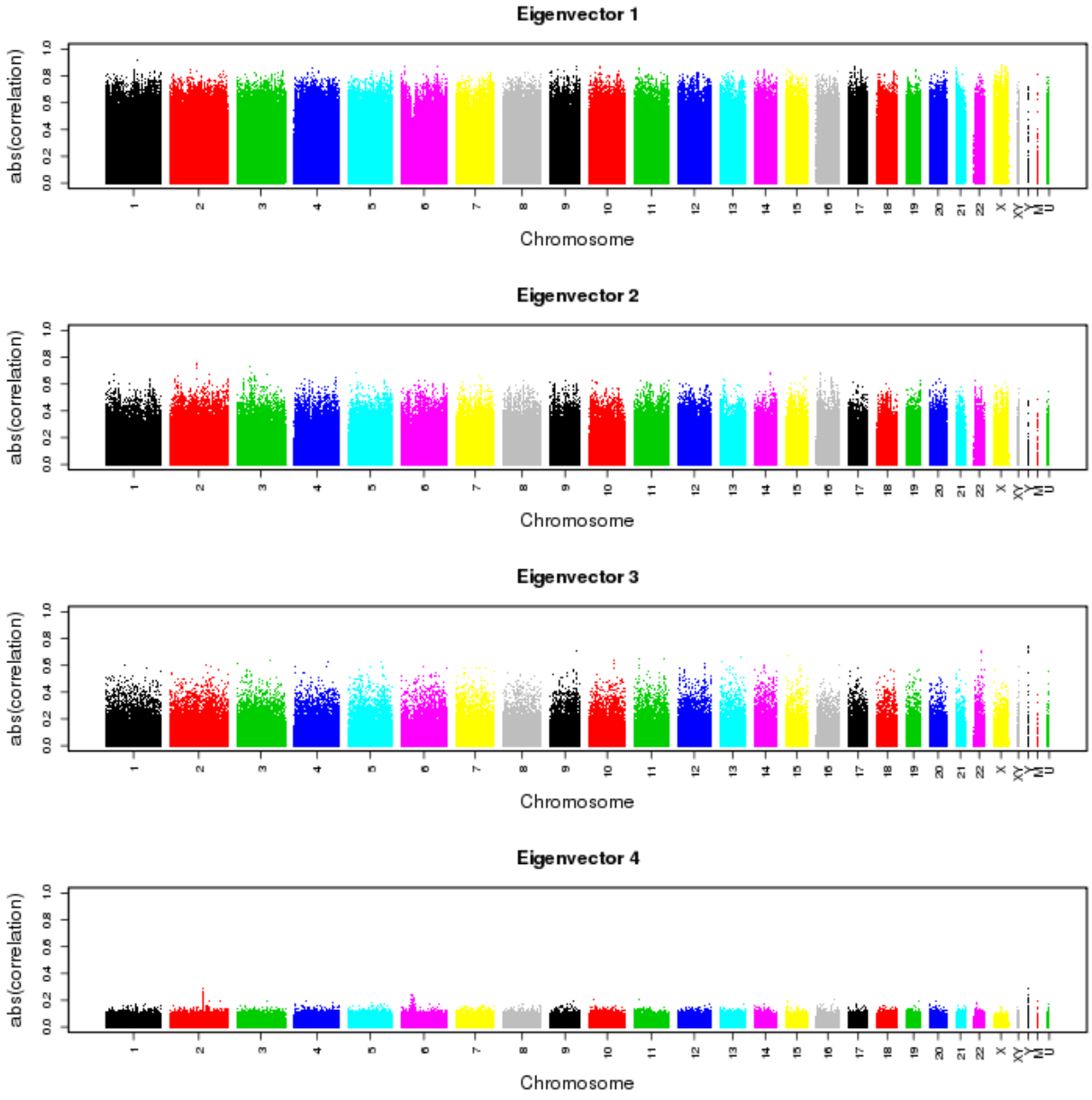



Figure 17: SNP position versus correlation between SNP genotype (0, 1 or 2) and each of the first 8 eigenvectors. These eigenvectors are from the PCA of all unrelated study subjects from Phase 3.

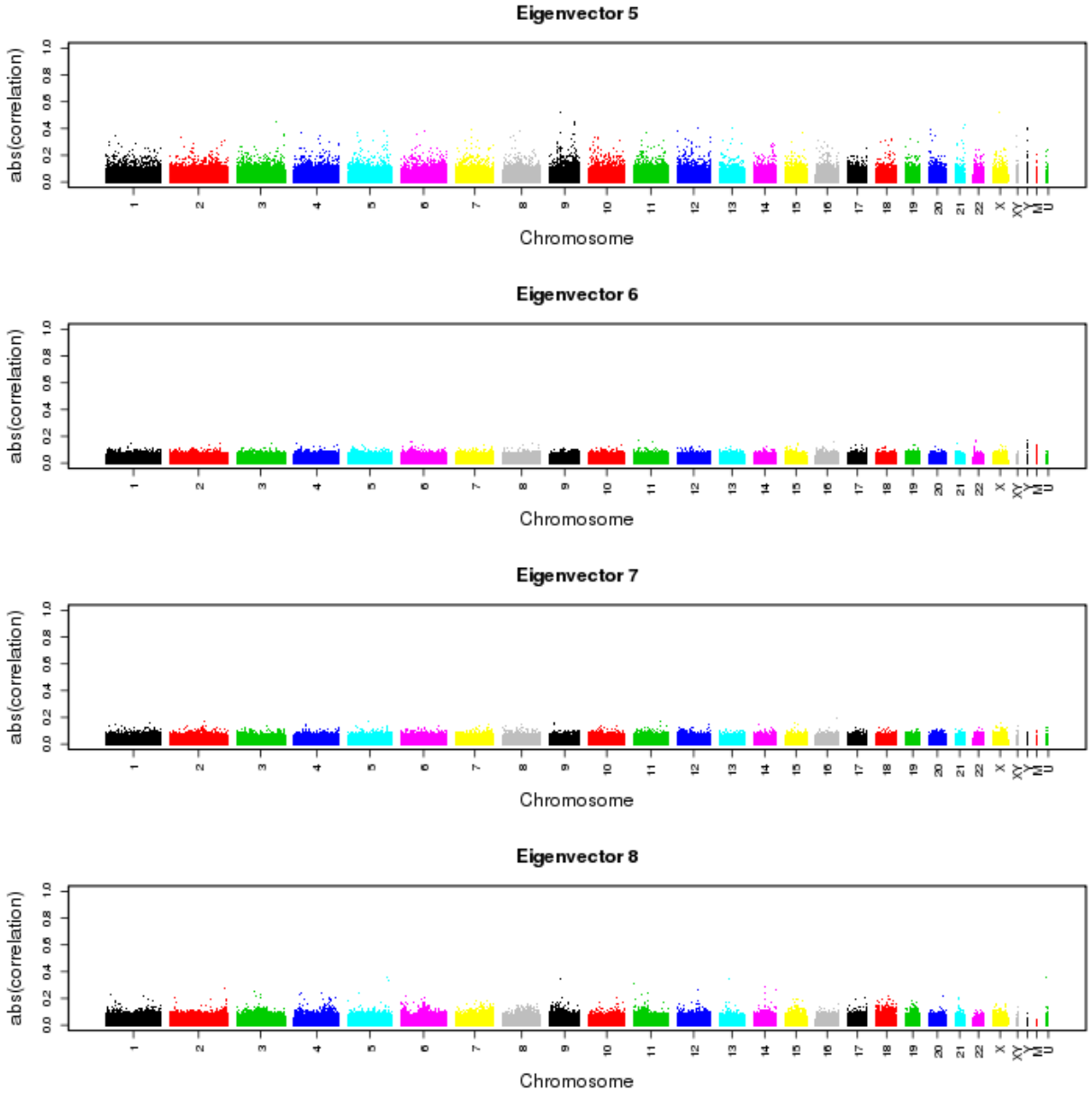



Figure 17: Continued.

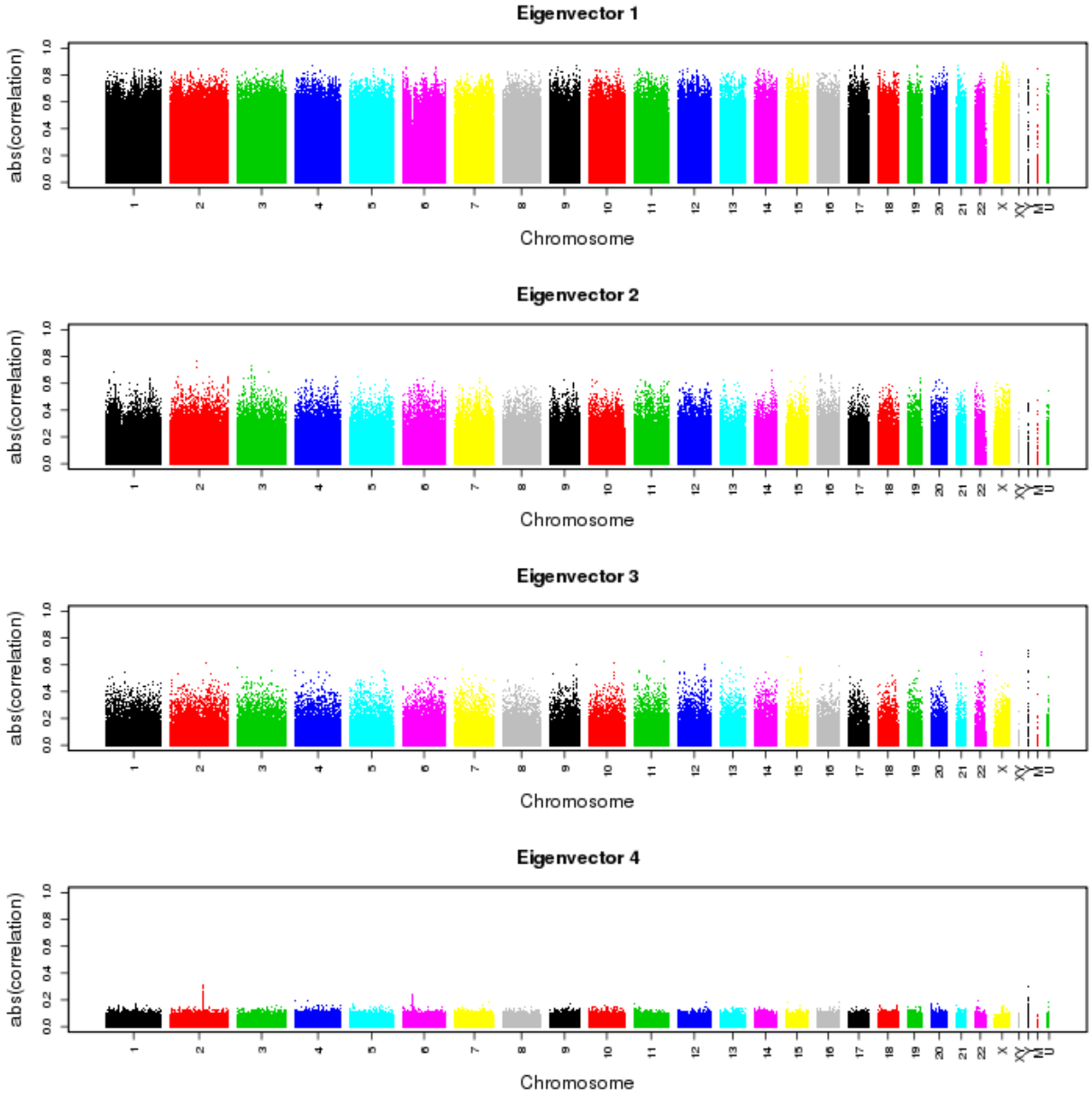



Figure 18: SNP position versus correlation between SNP genotype (0, 1 or 2) and each of the first 8 eigenvectors. These eigenvectors are from the PCA of all unrelated study subjects from combined Phases 1-2 and Phase 3.

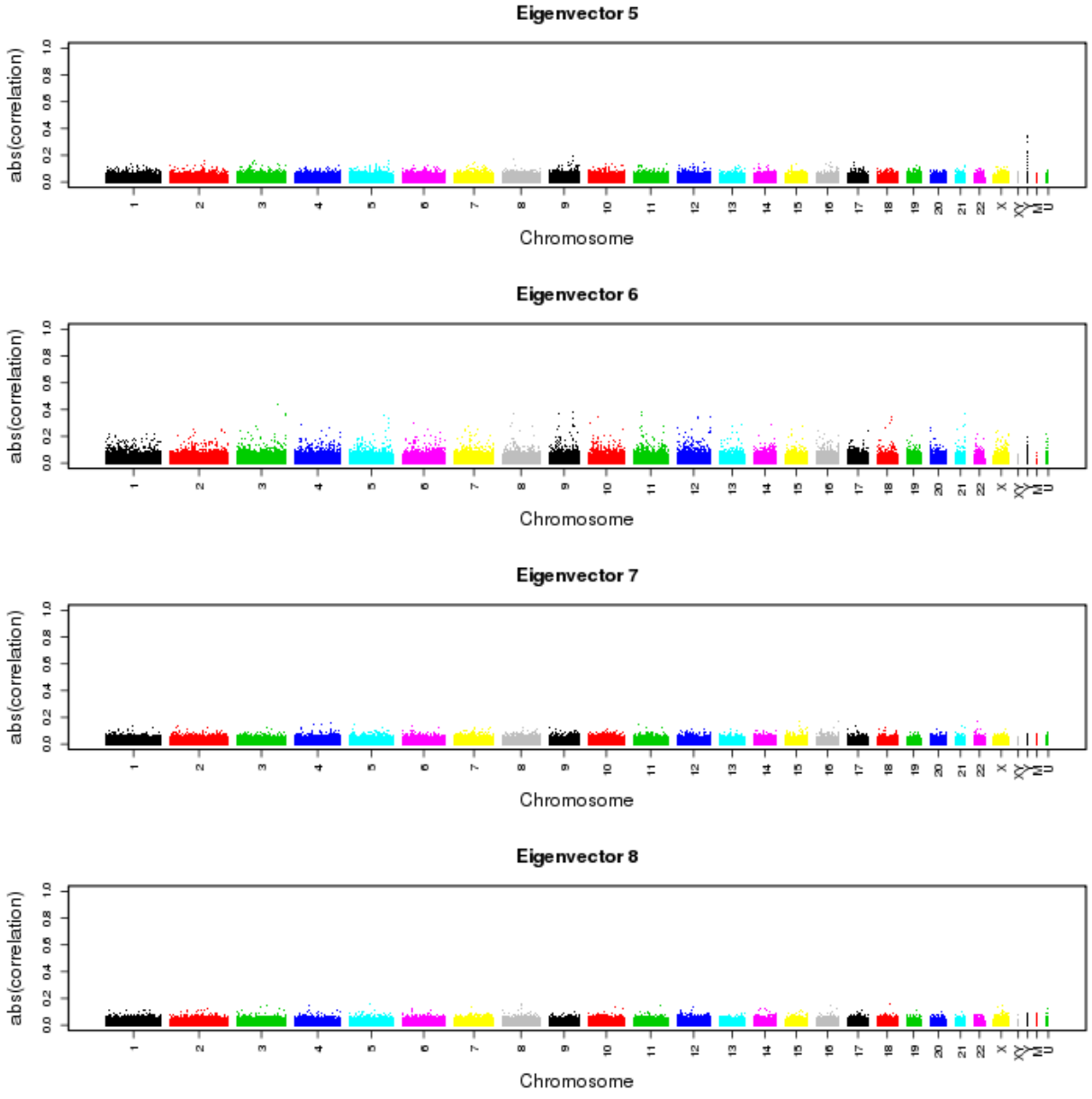



Figure 18: Continued.

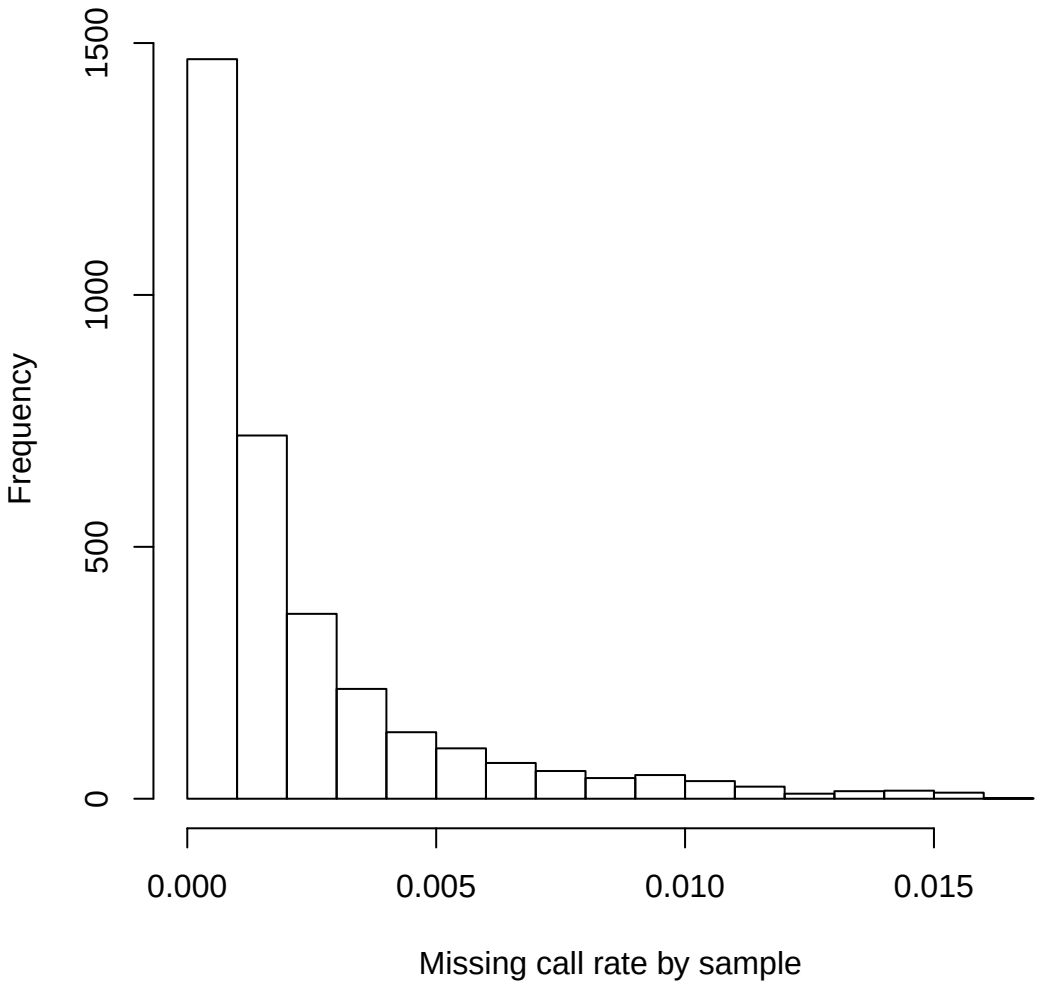



Figure 19: Histogram of the missing call rate per sample (*missing.e1*).

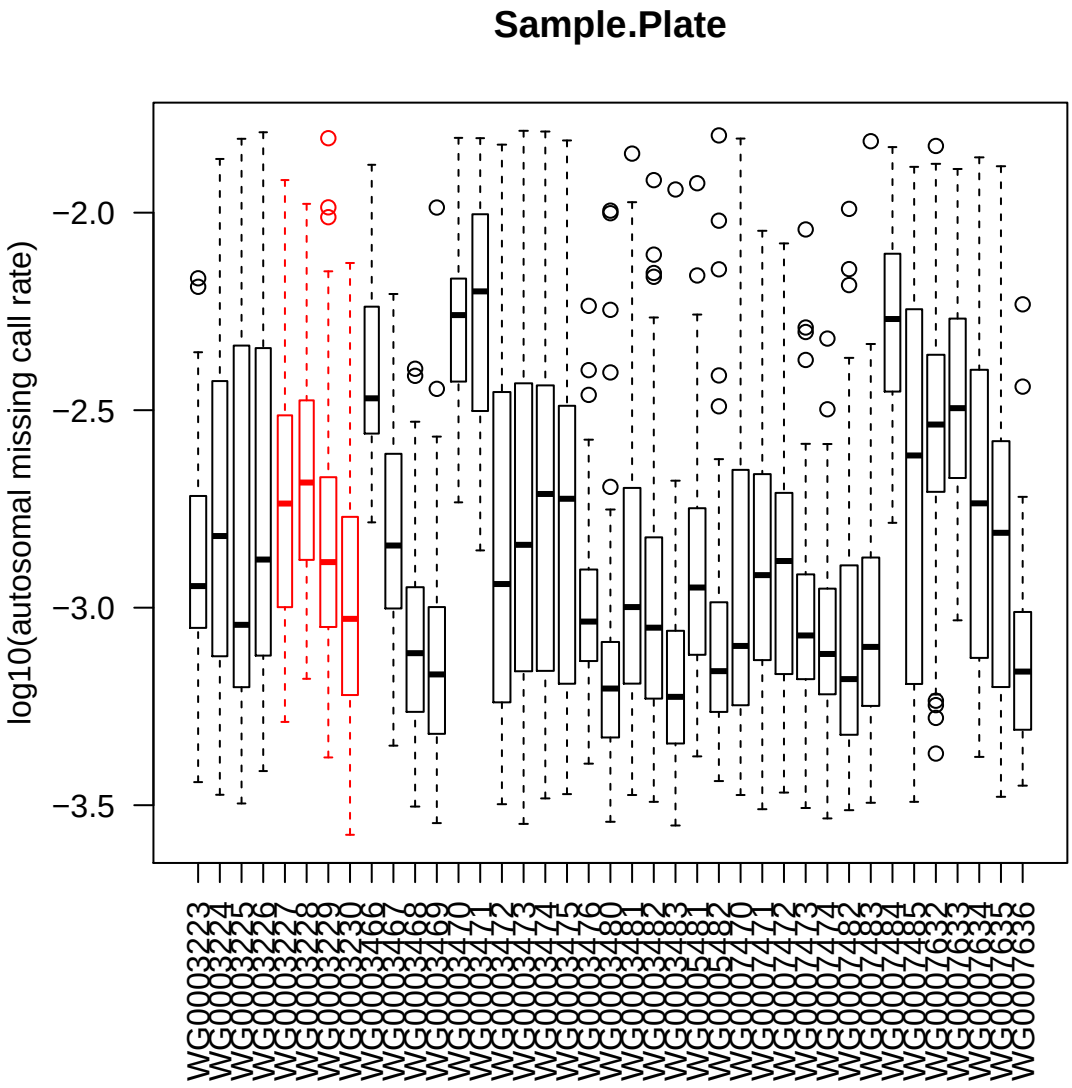



Figure 20: Boxplot of missing call rate for study samples categorized by genotyping plate. Red boxes indicate plates containing samples that failed in the first round of genotyping and were re-genotyped together.

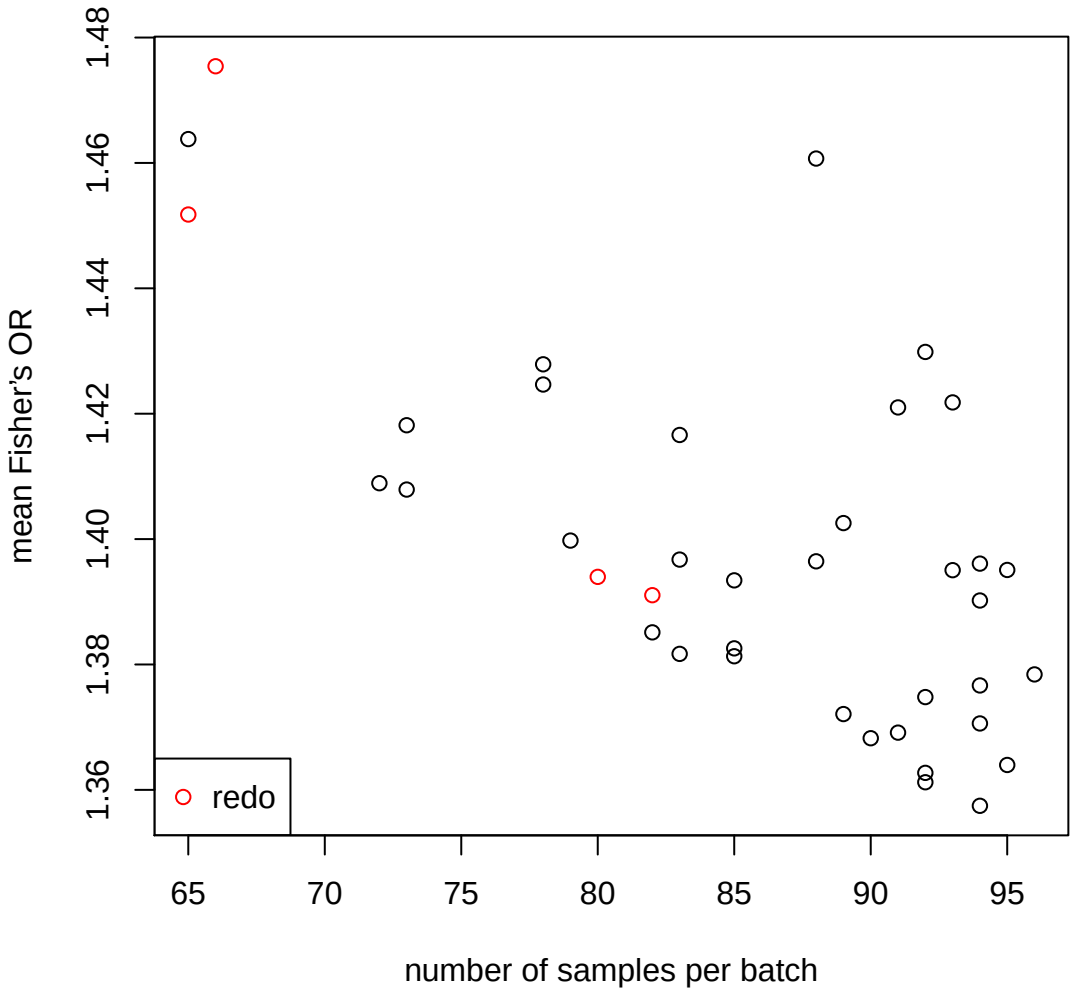
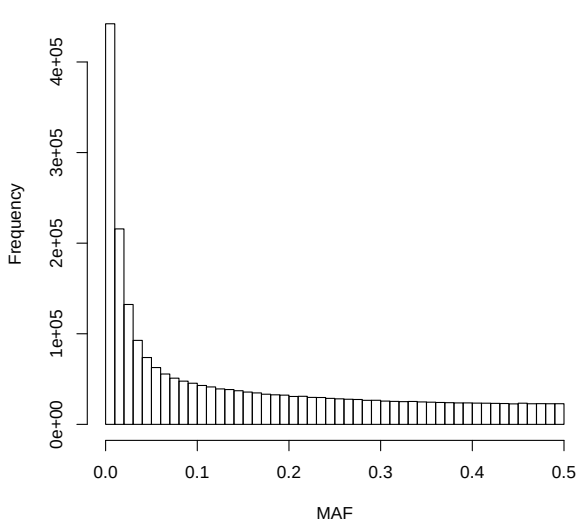
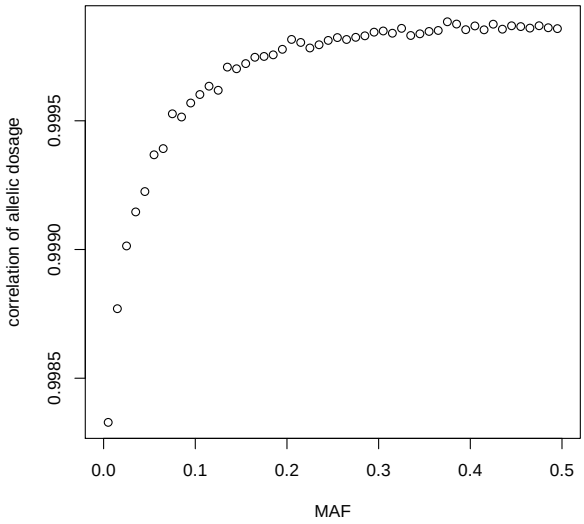


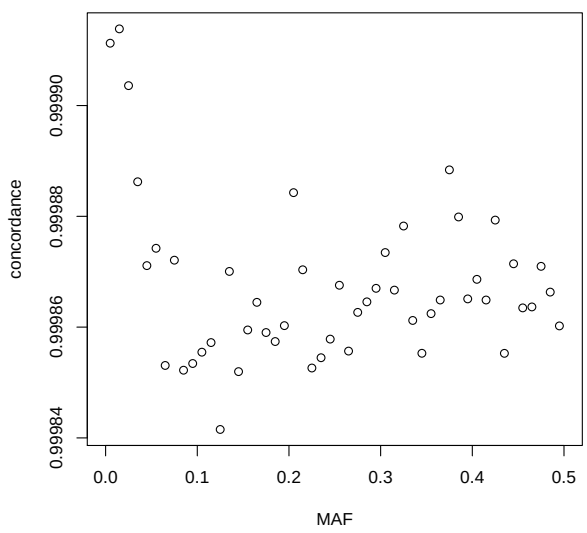

Figure 21: Mean odds ratio from Fisher's exact test of allele frequency plotted against fraction of self-identified 'White' samples per plate. The dashed vertical line is the mean over all plates.



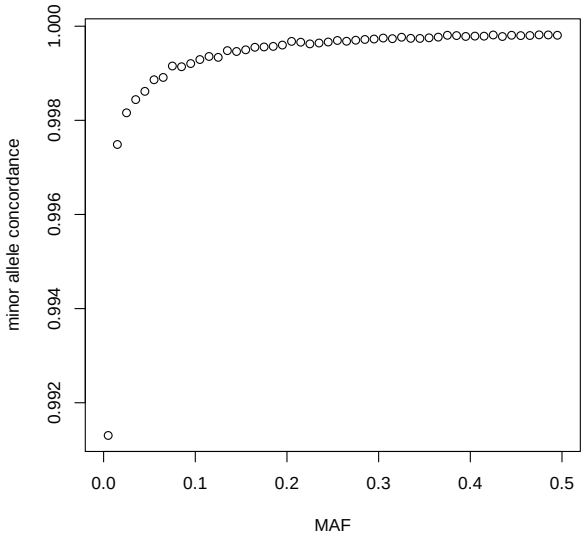

(a) Distribution of minor allele frequency.




(b) Correlation of allelic dosage.




(c) Overall concordance.




(d) Minor allele concordance.

Figure 22: Summary of concordance by SNP over 103 duplicate sample pairs, binned by minor allele frequency.

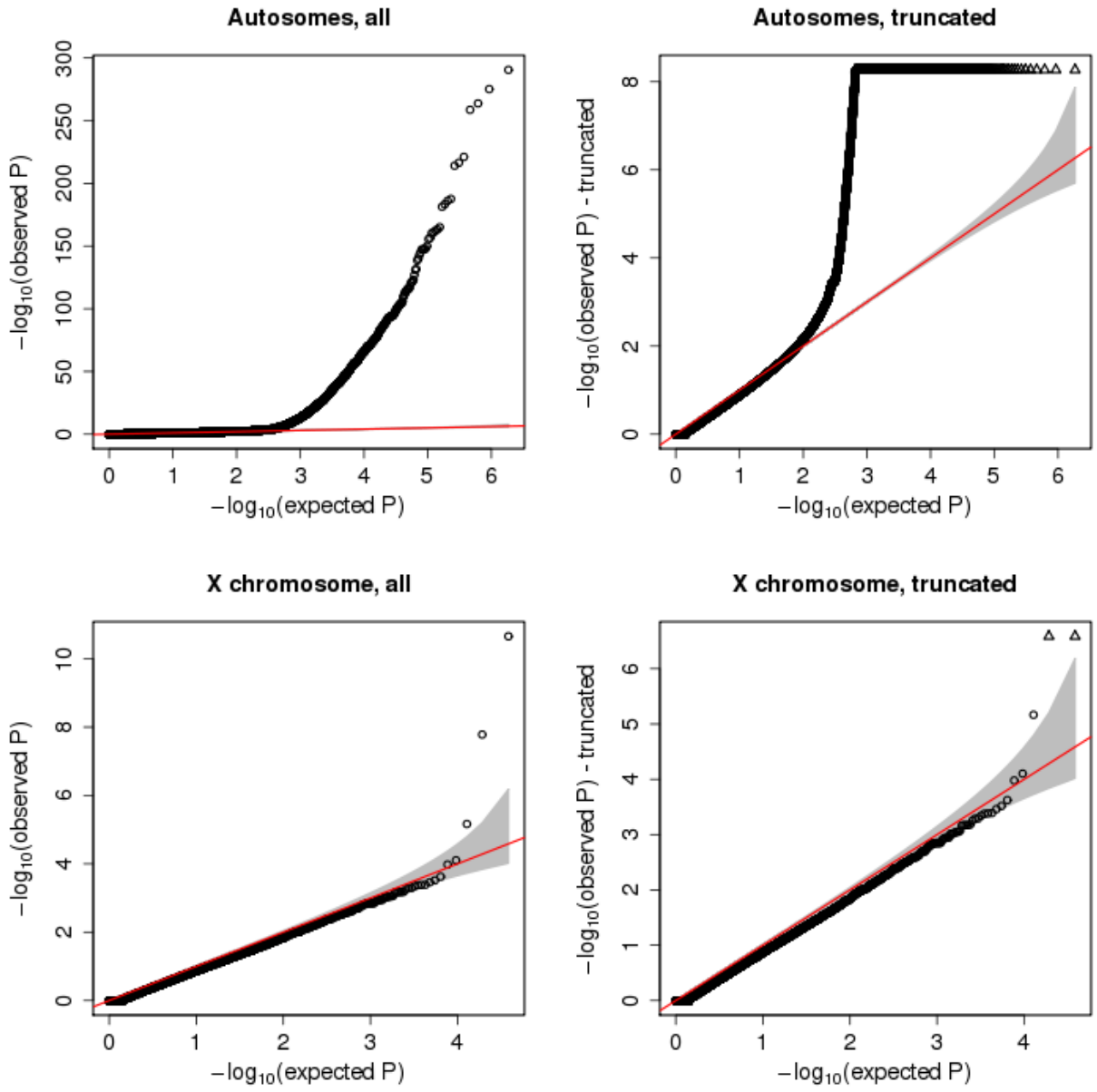



Figure 23: Quantile-quantile plots for $-log_{10}(p)$ from Fisher's exact test of Hardy-Weinberg equilibrium in European-ancestry subjects. Plots in the left column show all SNPs, whereas those in the right column have the Y-axis truncated to show more clearly the point of deviation from expectation.

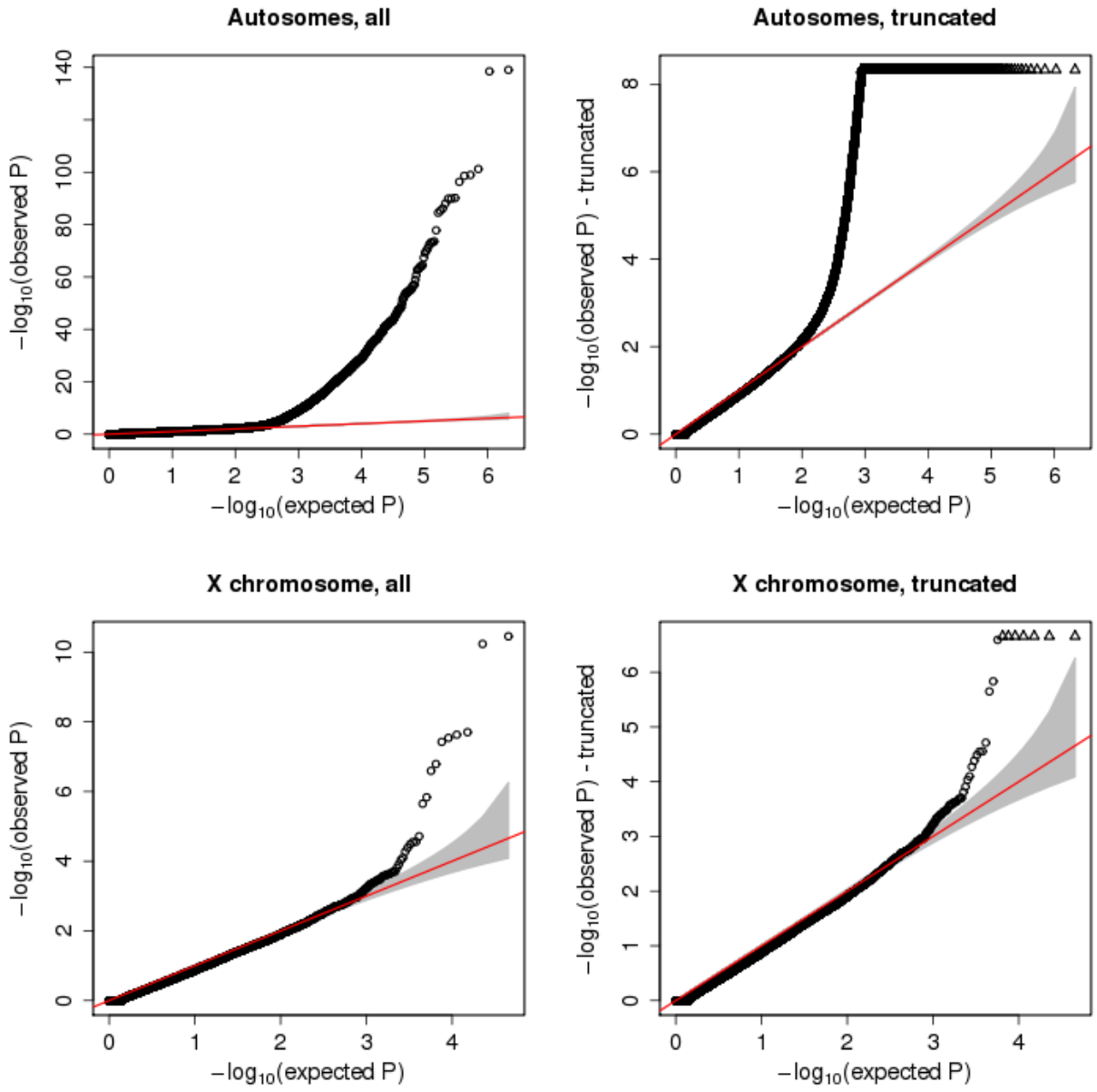
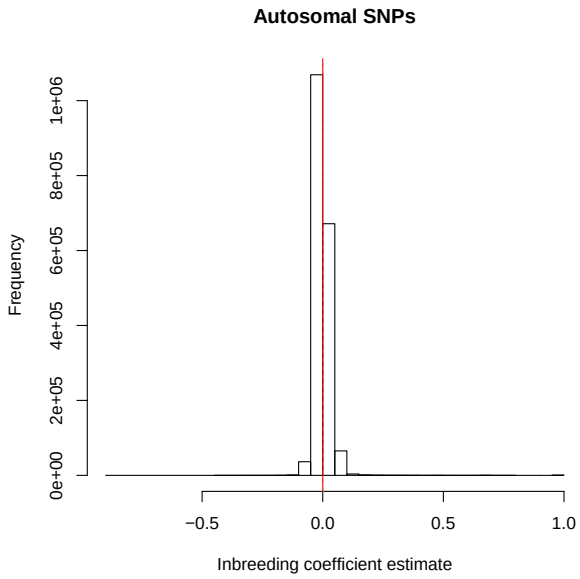
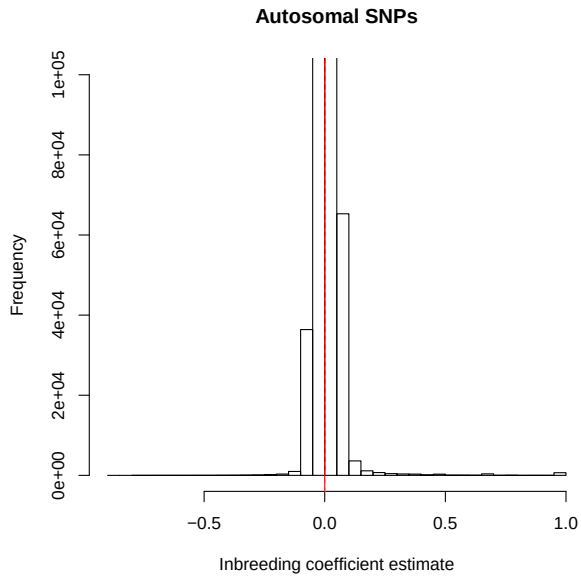


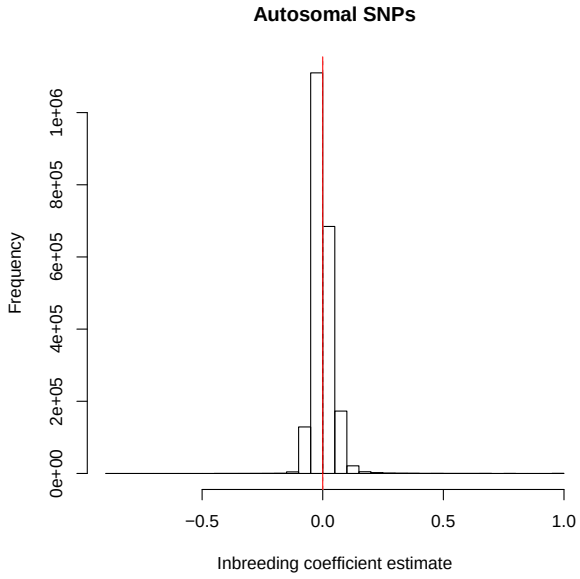

Figure 24: Quantile-quantile plots for $-log_{10}(p)$ from Fisher's exact test of Hardy-Weinberg equilibrium in African-ancestry subjects. Plots in the left column show all SNPs, whereas those in the right column have the Y-axis truncated to show more clearly the point of deviation from expectation.



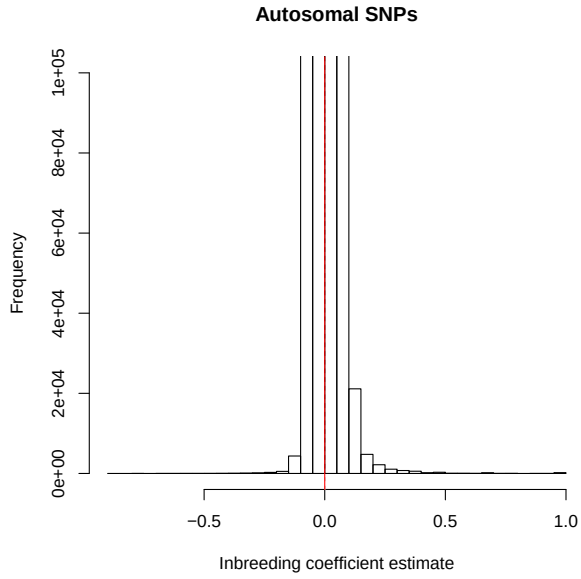

(a) European, all autosomal




(b) European, truncated

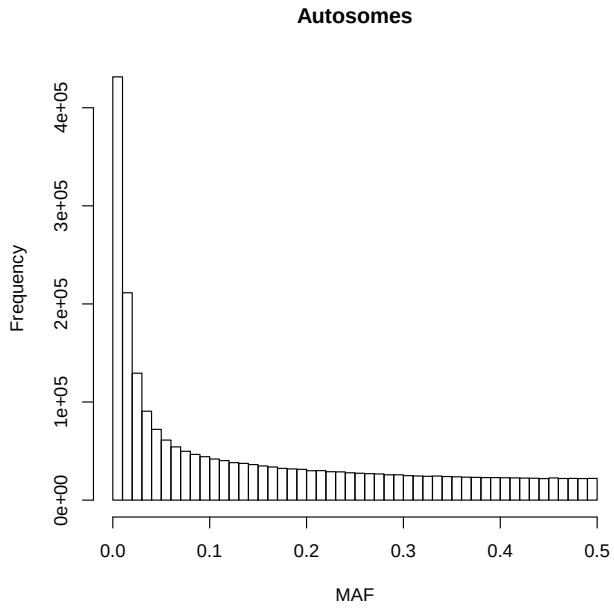



(c) African, all autosomal

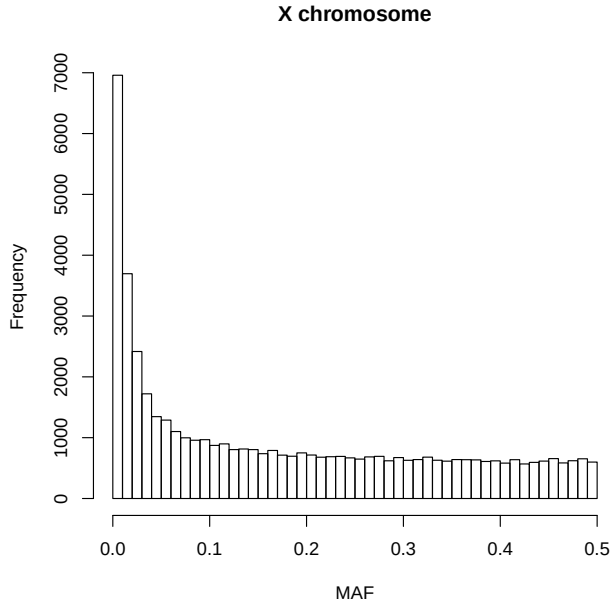



(d) African, truncated

Figure 25: Distribution of estimated inbreeding coefficient for all autosomal SNPs. The values range from -1 to 1.




(a) Autosomes




(b) X chromosome

Figure 26: Minor allele frequency distribution across all unrelated study subjects.

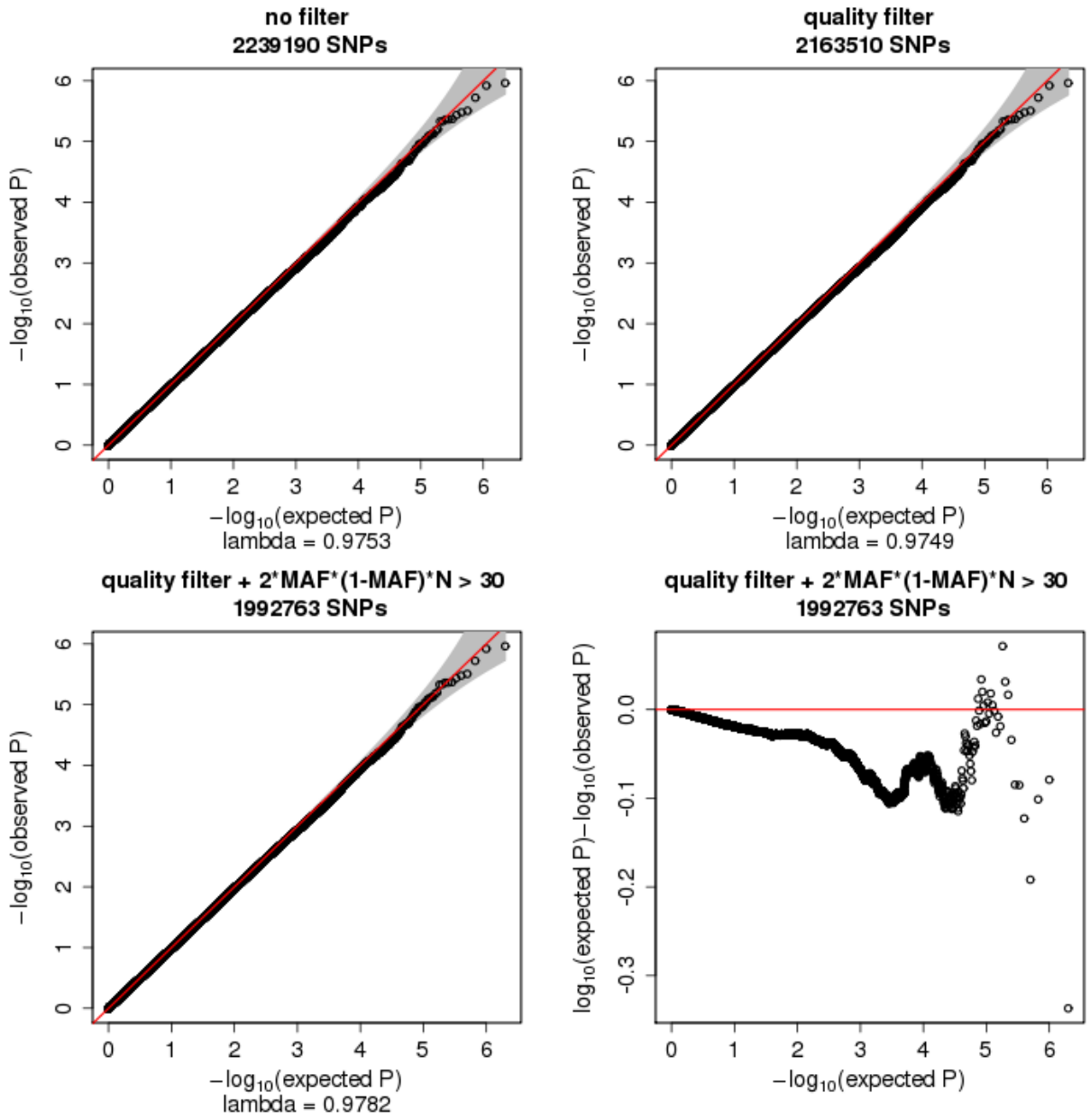



Figure 27: Quantile-quantile plots for preliminary association test using Phase 3 subjects only.

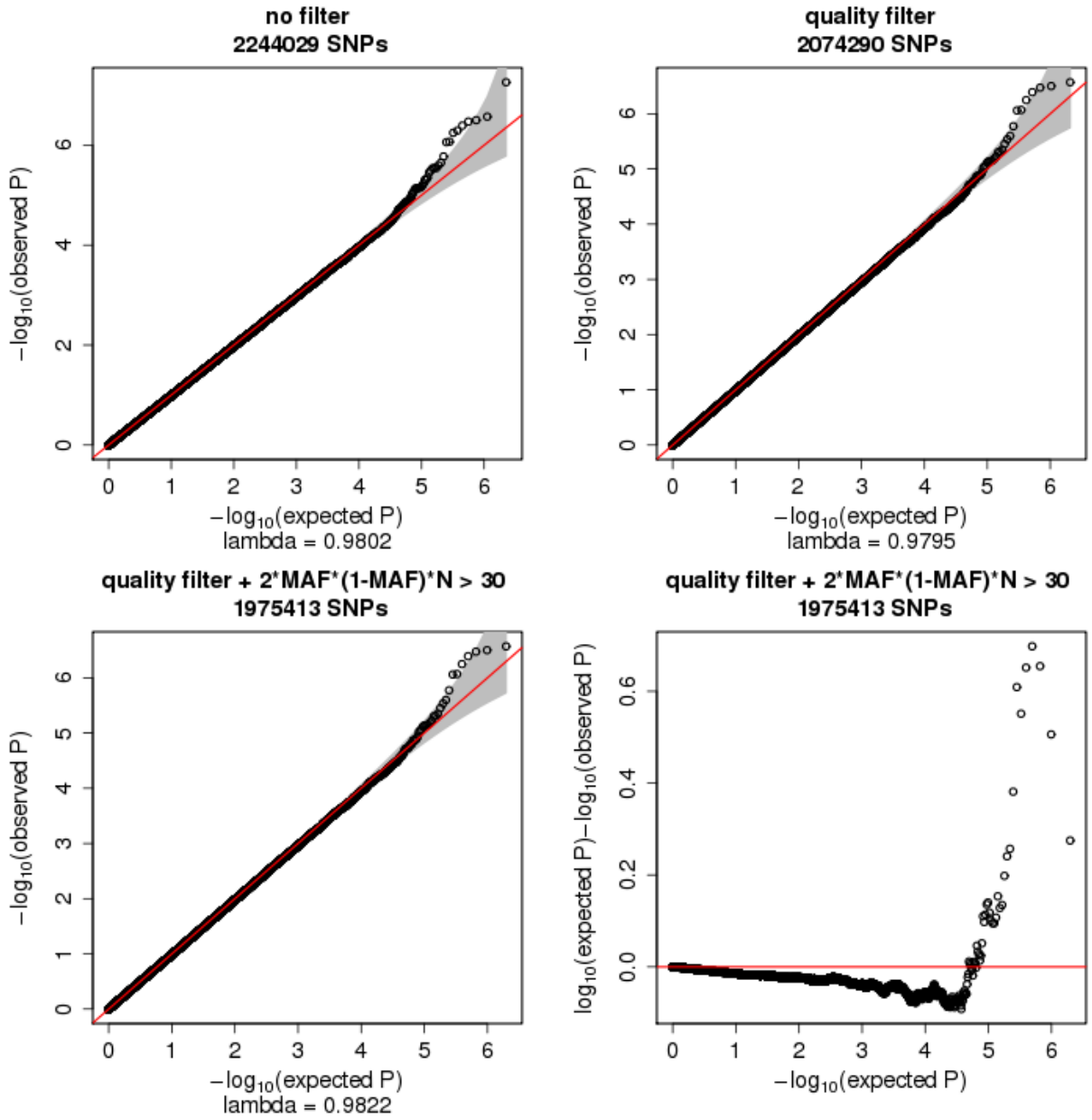



Figure 28: Quantile-quantile plots for preliminary association tests using Phases 1-2 and Phase 3 subjects combined.

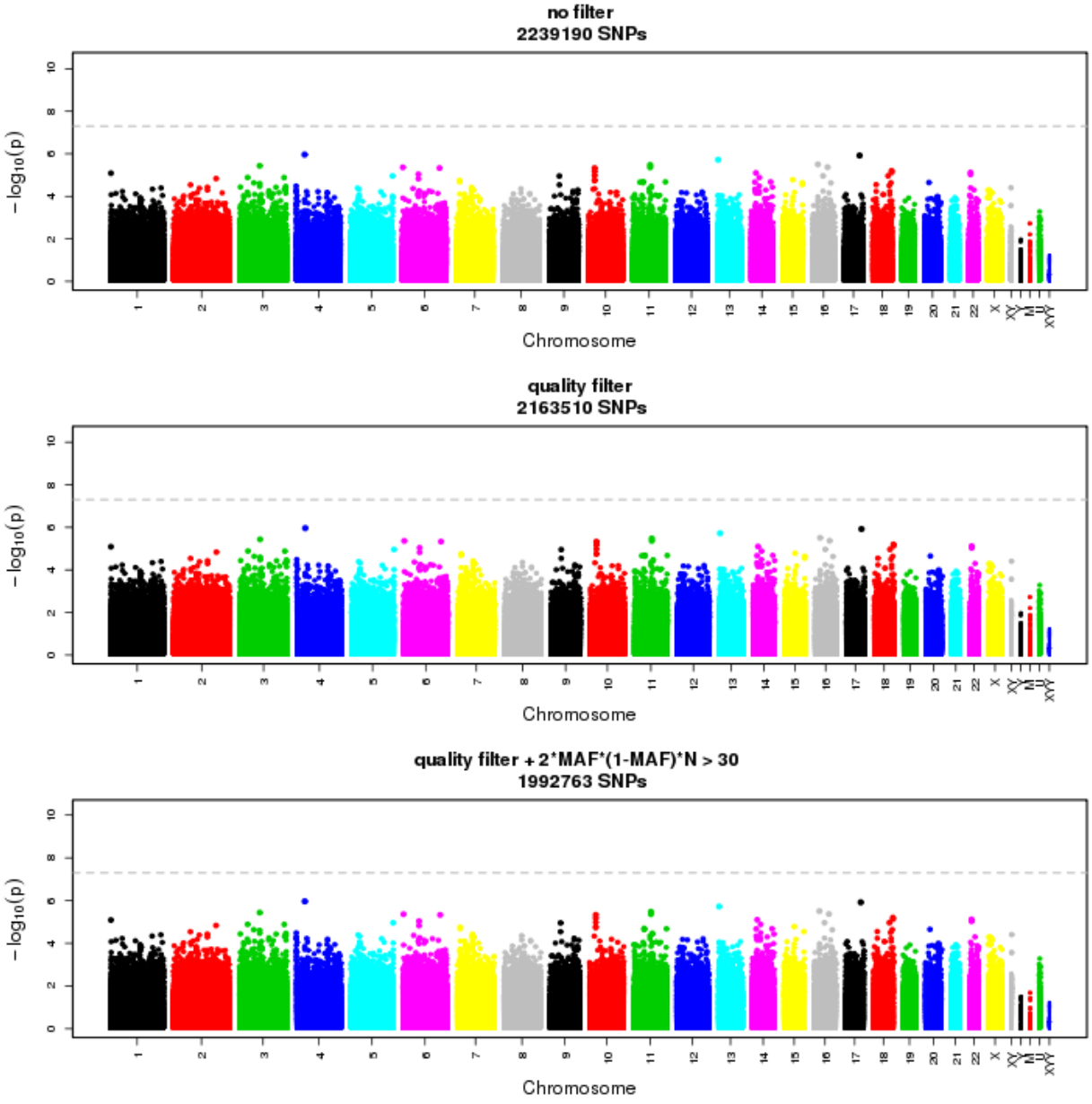



Figure 29: Manhattan plots for preliminary association test using Phase 3 subjects only.

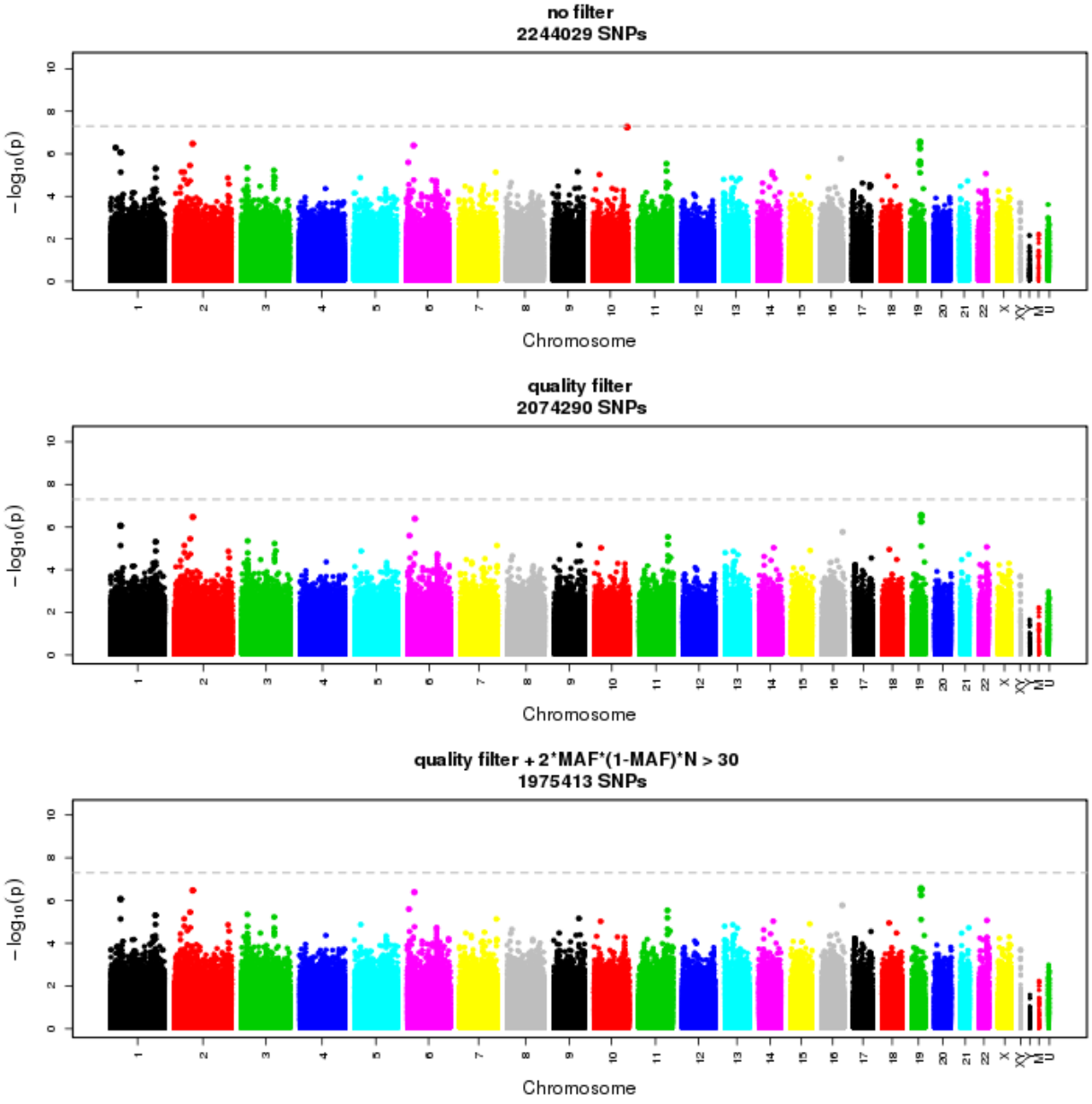



Figure 30: Manhattan plots for preliminary association tests using Phases 1-2 and Phase 3 subjects combined.

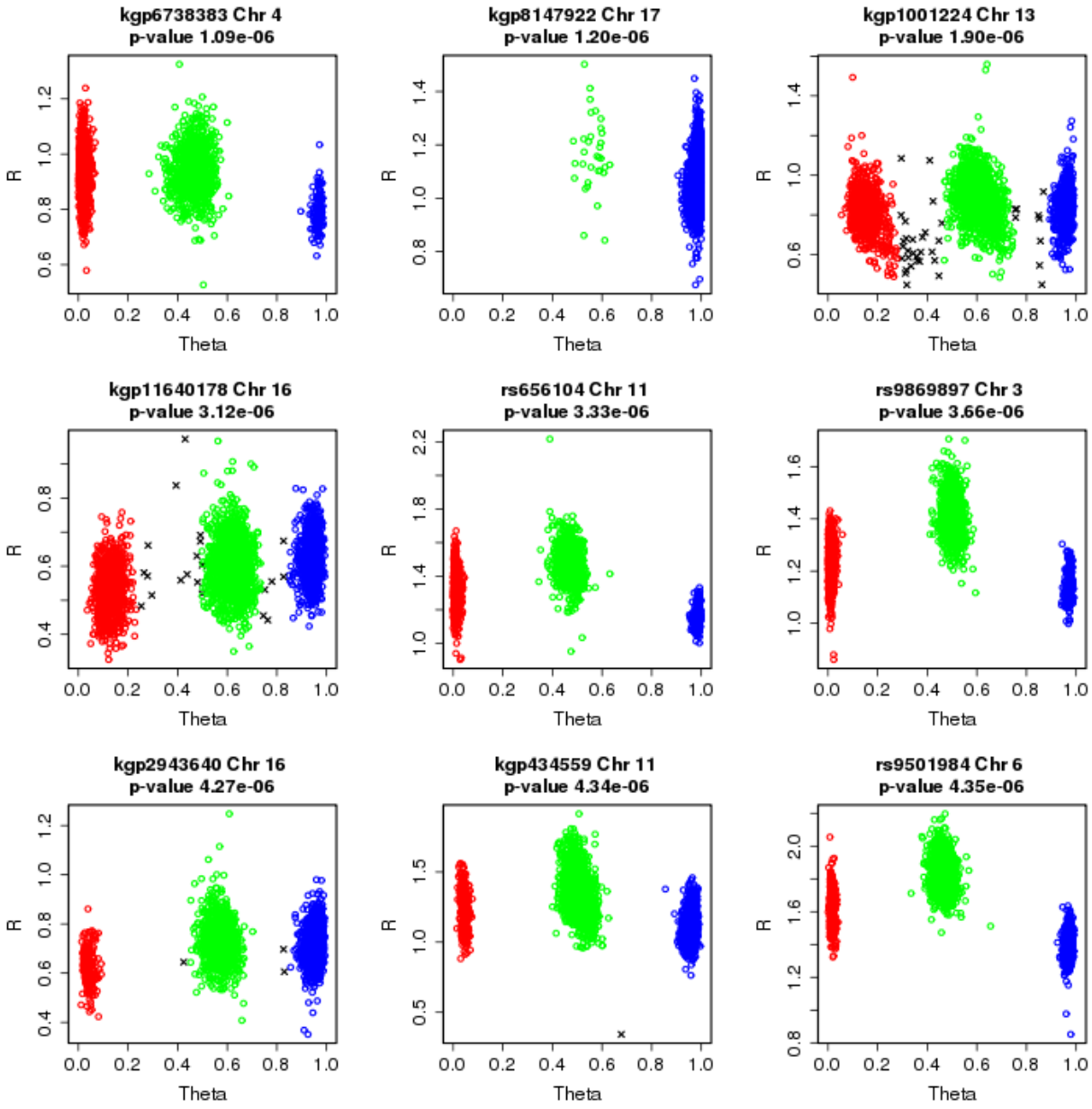



Figure 31: Genotype cluster plots for the top 9 SNPs from the preliminary association test after applying the quality and MAF ($> 2\%$) filters.