

Quality Control Report for Genotypic Data

University of Washington

October 24, 2016

Project: A Longitudinal Resource for Genetic Research in Behavioral and Health Sciences

Principal Investigator: Pamela Herd, PhD

Support: CIDR Contract # HHSN268201200008I

NIH Institute: National Institute on Aging

Contents

1	Summary and recommendations for dbGaP users	3
2	Project overview	3
3	Genotyping process	3
4	Quality control process and participants	4
5	Sample and participant number and composition	4
6	Annotated vs. genetic sex	4
7	Chromosomal anomalies	5
8	Relatedness	6
9	Population structure	7
10	Missing call rates	8
11	Batch effects	9
12	Duplicate sample discordance	9
13	Mendelian errors	10
14	Hardy-Weinberg equilibrium	10
15	Duplicate SNP probes	10
16	Sample exclusion and filtering summary	11
17	SNP filter summary	11

18	Minor allele frequency	11
19	Preliminary association tests	12
A	Project participants	13

List of Tables

1	Summary of recommended SNP filters	15
2	Summary of DNA samples and scans	15
3	Summary of subjects	15
4	IBD kinship coefficient expected values	16
5	P-values for eigenvectors	16
6	Summary of SNP missingness by chromosome	16
7	Duplicate sample discordance error rates and counts	17
8	MAF characteristics	17
9	Covariate selection	18

List of Figures

1	Sex check	19
2	Sex chromosome anomalies	20
3	Typical BAF scan	21
4	Anomalous BAF scan	22
5	Sex chromosome anomaly XX/X0, to filter	23
6	Sex chromosome anomaly XX/X0, not to filter	24
7	Sex chromosome anomaly XXY	25
8	Relatedness plot for all genotyped samples	26
9	PCA of study samples combined with HapMap 3	27
10	PCA of unrelated study samples without HapMap 3	28
11	PC-SNP correlation	29
12	PCA scree plot	31
13	PCA parallel coordinates plot	32
14	PCA of all study samples, direct and indirect	33
15	Histogram of missing call rate per sample	34
16	Histogram of missing call rate by sample batch	35
17	Boxplot of MCR categorized by plate	36
18	Boxplot of chromosome 1 intensity	37
19	Summary of concordance by SNP	38
20	Minor allele frequency distribution	39
21	PCA in samples for HWE testing	40
22	Distribution of estimated inbreeding coefficient	41
23	QQ plots of association test p-values	42
24	Manhattan plots of association test p-values	43
25	Genotype cluster plots	44

1 Summary and recommendations for dbGaP users

A total of 9,012 study participants were genotyped on the Illumina HumanOmniExpress array and included in this dbGaP posting. The median call rate is 99.95%, and the error rate estimated from 204 pairs of duplicated study samples is $2.84e-06$. Genotypic data are provided for all participants and SNPs. Generally, we recommend selective filtering of genotypic data prior to analysis to remove (1) large (> 5 Mb) chromosomal anomalies showing evidence of genotyping error and (2) whole samples with an overall missing call rate (MCR) $> 2\%$. In this study, all samples had $MCR < 2\%$. There were 35 large chromosome anomalies (31 in study samples) that were filtered, i.e. genotypes in the anomaly regions were set to missing. A composite SNP filter is provided, along with each of the component criterion so that users may vary thresholds (see Table 1). A preliminary association test is described as an example of how to apply the recommended filters. Additional specific recommendations for filtering genotype data are *italicized* in this report.

2 Project overview

The Wisconsin Longitudinal Study (WLS)¹ is a long-term study of a random sample of men and women who graduated from Wisconsin high schools in 1957 and their siblings. The WLS panel started out with a panel of 10,317 members from the class of 1957. Over time a second panel of 8,734 randomly selected siblings of the original graduate panel were recruited for the study. Of these combined panel members 9,027 contributed saliva for genetic analysis. Survey data were collected from the original respondents or their parents in 1957, 1964, 1975, 1992, 2004, and 2011 and from a selected sibling in 1977, 1994, 2005, and 2011. WLS data provide a detailed record of educational, social, psychological, economic and mental and physical health characteristics of a relatively homogeneous population that is almost entirely of Northern and Western European ancestry. Saliva was first collected in 2007-8 by mail. Additional samples were collected in the course of home interviews that began in March 2010. The addition of genetic data to WLS complements the store of extensive WLS phenotypic data and takes advantage of recent developments that have vastly increased opportunities for genetic studies of aging, behavior, cognition, personality, mental health, health, disease, and mortality. Researchers interested in linking the genetic data to the WLS survey data should email wls@ssc.wisc.edu.

3 Genotyping process

A total of 9,472 study samples, including planned duplicates, were successfully genotyped at the Center for Inherited Disease Research (CIDR) at Johns Hopkins University. These 9,472 study samples include 223 samples derived from two unique WLS control non-participants, which are not included in the dbGaP posting. There were 198 HapMap samples included as genotyping controls. Except for HapMap controls, all DNA samples were extracted from stored saliva using a modified Oragene extraction protocol, in order to amend the samples that were frozen prior to the extraction. Samples were genotyped in batches corresponding to 96-well plates with one batch per plate. On average, each batch contained two HapMap controls and three duplicate study samples. For 204 pairs of study sample duplicates, the two members of each pair were genotyped on separate plates.

Genotyping was performed at CIDR using the Illumina HumanOmniExpress array (humanomniexpress-24-v1-1, BPM annotation version A, genome build GRCh37/hg19) and using the calling algorithm GenomeStudio version 2011.1, Genotyping Module version 1.9.4, GenTrain Version 1.0. The array consists of a total of 713,014 SNPs. Two updates were made to the initial Illumina manifest. First, prior to genotype calling, CIDR corrected chromosome designations for numerous XY SNPs initially annotated as X or Y. These SNPs occur in pseudo-autosomal (PAR1, PAR2) regions or in the X-translocated region (XTR). Second, prior to genotype data cleaning, genomic positions were adjusted for insertion/deletion (indel) variants to match the convention used in the 1000 Genomes Project imputation

¹<http://www.ssc.wisc.edu/wlsresearch/>

reference panels. (See “chrom”, “chrom.ilm”, “position”, and “position.ilm” in “SNP_annotation.csv” for more details on chromosome and position updates.) While the array contains both SNPs and non-SNP variants (i.e., indels), in this report we use the term “SNP” more generically to refer to all genotyped variants.

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 UW GAC (University of Washington Genetic Analysis Center), the study investigator’s team, and to dbGaP. These data were analyzed by the analysis team at UW GAC, and the results were discussed with all groups in periodic conference calls. Key participants in this process and their institutional affiliations are given in Appendix A. The results presented here were generated with the R packages *GWASTools* [1], *GENESIS* [2], and *SNPRelate* [3] unless indicated otherwise. The methods of QA/QC used here are described by Laurie et al. [4].

5 Sample and participant number and composition

In the following description, 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 9,606 samples (including planned duplicates) from study subjects were put into genotyping production, of which 9,472 were successfully genotyped and passed CIDR’s QC process (Table 2). The subsequent QA/QC process did not identify any further sample exclusions due to low sample quality; however, 12 scans of questionable identity were identified and are excluded from posting. A further six samples were included in most analyses but later removed from the posted dataset due to withdrawn consent. The set of scans to be posted include 9,231 study participants and 198 HapMap controls. The 9,231 study scans derive from 9,012 unique subjects and include 219 pairs of duplicate scans (Table 3). Note one member of each of 15 monozygotic (MZ) twin pairs is counted as a duplicate scan.

The study subjects occur as 4,601 singletons and 2,263 families of 2–4 members each. Most study families consist of a priori known sibling pairs, though additional first, second, and third degree relationships were discovered by examining genetic relationships, discussed further in Section 8. The 198 HapMap control scans derive from 97 unique subjects, of which all are replicated two or more times. The HapMap controls include 12 singletons, two duos, and 24 trios, and one large CEU family².

6 Annotated vs. genetic sex

To compare annotated and genetic sex, we examine both X chromosome heterozygosity and the means of the intensities for X and Y chromosome probes. The expectation is that male and female samples will fall into distinct clusters that differ markedly for both metrics. The plots of X and Y chromosome intensity and heterozygosity in Figure 1 show the expected patterns: two distinct clusters corresponding to male and female samples. There were four discrepancies between annotated (expected) and genetic (observed) sex in this study. Two of the discrepancies could not be resolved or explained. These two samples are excluded from the dbGaP posting due to questionable identity. One discrepancy was explained by a known sex reassignment. The last discrepancy did not result in a sample exclusion because the sample showed the expected genetic relationship with a known full sibling and was thus assumed to have a correct linking between genetic and phenotypic records.

Deviations from expected intensity and heterozygosity on the X and Y chromosomes can also be used to detect potential sex chromosome anomalies. We observed the following potential sex chromosome anomalies: four XXY males, one XYY male, four XY/Y0 or “loss of Y” males, three XXX females, and six XX/X0 females, highlighted in Figure 2. Most of these likely sex chromosome anomalies were also identified by CIDR

²CEU: Utah residents with Northern and Western European ancestry from the CEPH collection

in their initial QC process. These samples were examined further by viewing BAF/LRR plots, discussed more 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) [5, 6]. 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) [7] 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 [8] for a full description and application of these methods. 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. To remove large (> 5 Mb) chromosomal anomalies that show evidence of genotyping error prior to analysis, we recommend selective filtering of genotypic data, effected by setting genotypes to missing in the anomaly region(s).

In this study, we detected 113 large (> 5 Mb) chromosomal anomalies in 98 unique subjects (94 excluding HapMap subjects), including 39 whole chromosome anomalies across both the autosomes and sex chromosomes. The file “chromosome_anomalies.csv” provides information on these 113 anomalies, including breakpoints. We recommend filtering genotypes in anomaly regions only when the anomaly appears to lead to an excess of miscalled (vs. just missing) genotypes. After reviewing BAF and LRR plots for the 113 anomalies in question, we recommend filtering 35 anomalies. This includes the following sex chromosome anomalies (see Section 6 and Figure 2): chromosomes X and XY for four XXY males, four XY/X0 males, and three of the six XX/X0 females (those with a split in the BAF band wide enough to cause genotyping errors). Note we do not recommend any X, Y, or XY filtering for the XXX females, XXY males, or XX/X0 females with narrow BAF band splits.

Next we present BAF and LRR plots for a subset of these anomalies, including a non-anomalous plot for comparison. Figure 3 shows BAF/LRR plots for chromosome 10 in Sample A. This chromosome shows a typical pattern, with LRR centered at 0 and three BAF bands at 0, 0.5, and 1, corresponding to AA, AB, and BB genotypes, respectively. Figure 4 shows BAF/LRR plots for chromosome 9 in the same sample (Sample A). This anomaly is a partial (25 MB) UPD recommended for filtering: the split in the heterozygous band is wide enough to result in many of the truly heterozygous genotypes miscalled as homozygous. The lack of change in LRR in the anomalous region is indicative of UPD, a copy-neutral anomaly, as compared to a partial deletion or duplication.

The next anomaly examples we present are sex chromosome anomalies first mentioned in Section 6 (see also Figure 2). Below “XY” SNPs refer to SNPs located in homologous regions of the X and Y chromosomes: the pseudoautosomal regions PAR1, PAR2, and XTR (X-translocated region). While karyotypically normal males are hemizygous for the non-pseudoautosomal portions of chromosome X and Y and thus have only one allele per SNP (i.e., A₋ or B₋ genotypes), for XY SNPs they should have two alleles (AA, AB, or BB genotypes).

Figure 5 shows BAF/LRR plots for one of the XX/X0 females whose chromosome X and XY genotypes are recommended for filtering. The split in the heterozygous band seen in the BAF plot is wide enough to result in many truly heterozygous genotypes miscalled as homozygous. For comparison, Figure 6 shows chromosome X for one of the XX/X0 female samples in which the BAF split is narrow enough not to require filtering. The narrower split suggests a higher ratio of XX:X0 cells, which does not lead to miscalled

genotypes.

Figure 7 shows BAF/LRR plots on chromosome X for one of the four males with an apparent XXY karyotype. This plot is shaded to distinguish SNPs in the pseudoautosomal regions (in PAR1, PAR2, or XTR, plotted as green points) from the non-pseudoautosomal regions of chromosome X (SNPs plotted as fuschia points), rather than to distinguish different genotype calls. The plots are consistent with an XXY karyotype in that there is a heterozygous band at BAF = 0.5 in the non pseudo-autosomal regions, indicating two copies of the X chromosome, and a total of four BAF bands in the pseudo-autosomal regions indicating copy number of three (AAA, AAB, ABB, and BBB genotypes produced by the presence of two X chromosomes and one Y).

In addition to their use in detecting chromosome anomalies, we also examine BAF density plots and BAF/LRR plots for evidence of sample contamination (more than three 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.

8 Relatedness

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

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

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.

IBD coefficients were estimated using 94,261 autosomal SNPs and the KING-robust procedure [9], implemented in the R package *SNPRelate* [3]. The SNPs were selected by linkage disequilibrium (LD) pruning from an initial pool consisting of all autosomal SNPs with a MCR < 2% and minor allele frequency (MAF) > 5%, with all pairs of SNPs having $r^2 < 0.1$ in a sliding 10 Mb window. KING-robust was used as it is robust to population structure. KING-robust provides estimates of KC and IBS0 (the fraction of SNPs that share no alleles), from which different relationship types can be inferred.

We used the KING-robust estimates to assess consistency between expected and observed relationships, including duplicate sample pairs. Of the 9,045 unique genotyped study participants, 4,500 were intially annotated as belonging to one of 2,250 pedigrees. Following from the sibling study design of WLS, expected relationships included full siblings, half siblings, and MZ twin pairs. The remaining 4,545 samples were expected to be mutually unrelated to each other and to any participants in the known pedigrees. Additionally, there were 204 pairs of expected duplicate study samples and numerous parent-offspring (PO) relationships expected among HapMap genotyping controls.

We used the observed genetic relationships among samples to verify the a priori expected relationships described above. Figure 8 shows the KC and IBS0 estimates for 15,917 pairs of samples with $KC > 1/32$, corresponding to a fourth or lower degree relationship. In this plot, symbol shape denotes expected (circles) versus unexpected (triangle) relationships, and symbol color denotes expected relationship type, prior to implementing pedigree resolutions (described below). All planned study duplicates and HapMap PO relationships were observed as expected. However, there were numerous discrepancies between expected and observed relationships among study subjects, including unexpected but observed and expected but unobserved relationships.

The study investigator group was able to reconcile the majority of these discrepancies by reviewing prior survey and other participant records and adjusting the pedigree structure accordingly. The pedigree resolution process involved identifying additional sibling and MZ twin pairs and joining previously unconnected families. Several second (e.g., half siblings or avuncular) and third degree (e.g., first cousin) relationship pairs could not be disambiguated into a particular pedigree structure.

The pedigree resolution process also uncovered some sample identity issues. First, two sample swaps affecting four samples were identified through sets of expected but unobserved sibling pairs. The sample swaps were resolved by switching the sample-level data (i.e., genetic data) for the given subject-level data (i.e., phenotype and pedigree information) within each sample swap pair. Second, a total of 10 samples were involved in expected but unobserved full sibling relationships, with no available explanation for the observed discrepancy. These 10 samples are excluded from the dbGaP posting due to questionable identity, in particular the potentially incorrect linking between genetic data and participant consent.

Per WLS Institutional Review Board stipulation, no pedigree file is provided for this study. Additionally, in sample annotation and PLINK genotype data files, family and parental identifiers do not reflect pedigree structure (i.e., family ID is set equal to subject-level ID, and parental IDs all equal 0). However, users can account for sample relatedness in downstream analyses using the IBD coefficient estimates provided in the file “Kinship_coefficient_table.csv.” *For an analysis that assumes all participants are unrelated, we recommend selecting a maximal set of unrelated subjects using the “unrelated” flag in “Sample_analysis.csv.” Note that graduates (versus their siblings) were preferentially selected into the maximal unrelated set.*

9 Population structure

To investigate population structure, we use principal components analysis (PCA), essentially as described by Patterson et al. [10] but implemented in the R package *SNPRelate* [3]. The motivation for running PCA is twofold: (1) to identify homogeneous genetic ancestry groups for subsequent QA/QC steps and (2) to provide sample eigenvectors that can be used to adjust for population stratification in downstream analyses.

We and others have shown that it is often necessary to prune SNPs based on LD and other criteria prior to performing PCA. Pruning is done to avoid generating sample eigenvectors that are determined by small clusters of SNPs at specific locations, such as the LCT, HLA, or polymorphic inversion regions [11]. Therefore, the SNPs used in the PCA described below were selected by LD pruning from an initial pool consisting of all autosomal SNPs with a MCR < 2% and MAF > 5%. The LD pruning process selects SNPs from the initial pool with all pairs having $r^2 < 0.1$ in a sliding 10 Mb window. In addition, the 2q21 (LCT), HLA, 8p23, and 17q21.31 regions were excluded from the initial pool of eligible SNPs.

We performed two PCA: the first PCA (A) combined unique (non-duplicated) study samples with an external set of HapMap population controls to establish the ancestry orientation and to identify possible population group outliers. The second PCA (B) was performed on a set of unique study subjects unrelated at the third degree level (6,543 samples). LD pruning was done separately for the two PCA, using sample sets described below. Additionally, we used the study-only PCA, run using the maximal unrelated sample set, to project sample eigenvectors into relatives, described in part (C). Note these PCA included six samples that were ultimately removed from the dbGaP posting due to revoked consent.

(A) **Combined PCA.** We performed PCA on non-duplicated study samples along with an external set of HapMap 3 [12] samples genotyped on the Illumina 1M array. A total of 9,018 study samples and 1,138 HapMap samples were included in this PCA. In addition to the general pruning process described above, the initial pool of SNPs included only those on both arrays with no genotyping discordances in the HapMap samples common to both datasets. LD pruning was performed in a set of unduplicated study samples unrelated at the third degree (6,543 samples), yielding 84,740 pruned SNPs.

Figure 9 shows the first two eigenvectors from the combined PCA. Samples are color-coded by either population, for HapMap samples³, or self-identified race group, for study samples. Most study samples (> 85%) self-identify as “white” and indeed the majority of study samples cluster with European ancestry HapMap populations. There are also some study samples with high proportions of Asian or African ancestry. Study samples of “unknown” self-reported race (13% of study samples) are mostly located in the European

³ASW: African ancestry in Southwest USA; CEU: Utah residents with Northern and Western European ancestry from the CEPH collection; CHB: Han Chinese in Beijing, China; CHD: Chinese in Metropolitan Denver, Colorado; GIH: Gujarati Indians in Houston, Texas; JPT: Japanese in Tokyo, Japan; LWK: Luhya in Webuye, Kenya; MKK: Maasai in Kinyawa, Kenya; MXL: Mexican ancestry in Los Angeles, California; TSI: Toscani in Italia; YRI: Yoruba in Ibadan, Nigeria

cluster, though some extend along either the EV2 arm, towards Asian samples, or along the EV1 arm, overlaid with self-identified African-American samples.

(B) **Study-only PCA.** We then performed PCA on a set of 6,543 unrelated study samples. LD pruning was performed in this same set of unrelated samples and yielded 93,315 pruned SNPs. Unrelated study samples were selected as described in Section 8 and can be identified with the “unrelated” variable in “Sample_analysis.csv.” Figure 10 shows the first two eigenvectors from this study-only analysis.

We examine plots of the correlation of each genotyped SNP with each eigenvector to determine whether the LD-pruning effectively prevented the occurrence of small clusters of SNPs that are highly correlated with a specific eigenvector. These plots are similar to GWAS “Manhattan” plots except the SNP-eigenvector correlation is plotted on the Y-axis, rather than an association test p-value. Figure 11 shows these plots for the first 8 eigenvectors from the PCA of all unrelated study participants. For the most part, no clusters of highly correlated SNPs are evident in these plots, indicating that each eigenvector is related to many SNPs distributed across all chromosomes.

The results from the study-only PCA were used for subsequent QA/QC steps. First, the PCA was used to select a homogeneous set of study samples for evaluating Hardy-Weinberg equilibrium as part of a SNP quality filter (see Section 14). Second, the PCA results were used to select sample eigenvectors for preliminary association tests. To determine which eigenvectors might be useful covariates to adjust for population stratification in downstream association testing, we examine the scree plot for the PCA (Figure 12), the parallel coordinates plot (Figure 13) and the association of each eigenvector with phenotype (here, the summary cognitive score “theta,” see Table 5). The scree plot shows the fraction of variance accounted for decreases markedly with each of the first six components and plateaus starting with the seventh component. The parallel coordinates plot is color-coded by self-identified race, where each vertical line represents an eigenvector and each piece-wise line between the vertical lines traces eigenvector values for a given subject. Regression analysis of the phenotype on the first twelve eigenvectors sequentially indicates significant association for EV1 and EV6, after correcting for multiple testing. Thus we decided to use EV1 - EV6 as covariates in association testing. Further discussion of model selection is in Section 19.

(C) **Projection into relatives.** In order to obtain eigenvectors for all study samples, including relatives of the unrelated set, we implemented the approach described by Zhu *et al.* [13]. In this approach, a maximal set of unrelated samples is analyzed to obtain SNP eigenvectors, which are then used to calculate sample eigenvectors for (i.e. project into) the remaining samples. Figure 14 is a plot of the first two eigenvectors from this PCA projection. The left panel shows direct sample eigenvectors while the right panel shows the indirect, or projected, sample eigenvectors. These plots illustrate that the indirect sample eigenvectors largely overlap the directly calculated sample eigenvectors. Results from this PCA analysis are given in the file “Principal_components.csv,” with the variable “type” indicating if the values were obtained through direct or indirect calculation. *Note that the self-identified race variable is not included in the dbGaP posting.*

10 Missing call rates

Two missing call rates are provided for each sample and for each SNP in the files “SNP_analysis.csv” and “Sample_analysis.csv” on dbGaP. The rates are calculated as follows: (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 0.000472 (*missing.e1*) and 0.000406 (*missing.e2*). Figure 15 shows the distribution of *missing.e1*. Generally we recommend filtering samples with MCR > 2%; however, in this study, all samples have MCR less than 2%. Thus no samples are recommended for exclusion based on MCR. The two missing call rates by SNP are identical. Table 6 gives a summary of SNP genotyping failures and missingness by chromosome type. For SNPs that passed the genotyping center QC (i.e., non-technical failure SNPs), the mean and median values of *missing.n1* are 0.00117 and 0.000311, respectively, and 99.1% of SNPs have a missing call rate < 2%. *We recommend*

filtering out samples with a MCR > 2% (although there are none in this study) and SNPs with a MCR > 2%.

An association between MCR and phenotype can lead to spurious associations, because missingness is often nonrandom [14]. We tested for a such a difference using linear regression of $\log_{10}(\text{miss.e1.auto})$ (autosomal missing call rate) on the summary cognitive score “theta.” We found a modest association between theta and MCR ($p = 0.037$). We recommend examining genotype cluster plots for any SNPs found to be significant in downstream association analysis, to rule out an artifactual association driven by non-random missingness.

11 Batch effects

Samples were processed together in the genotyping laboratory in batches consisting of 105 complete or partial 96-well plates. Some samples that failed the first round of genotyping were re-genotyped together on “redo” plates. There were 300 total samples across five such redo plates.

To identify any batch or plate effects, we plotted \log_{10} of the autosomal missing call rate (Figure 17) and mean Chromosome 1 intensity (Figure 18) for each plate. While there is a highly significant variation in MCR between different plates, all plates have a low mean MCR (see Figure 17). The mean of the average by-plate MCR is 0.00119. There are no outlier plates, as initially failing samples genotyped on the two redo plates are likely to have higher MCR. Chromosome 1 intensity profiles are also similar across plates. Ultimately we concluded that there are no problematic plate effects.

12 Duplicate sample discordance

Genotyping error rates can be estimated from discordance rates among duplicate sample pairs. The genotype at any bi-allelic SNP may either be called correctly or miscalled as either of the other two genotypes. If the two error rates for a duplicated genotyping instance of the same participant are α and β , respectively, then the probability of 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 study, because the median discordance rate over all sample pairs is $5.67\text{e-}06$, a rough estimate of the mean error rate is $2.835\text{e-}06$ errors per SNP per sample, indicating a high level of reproducibility.

In addition to estimating overall genotyping error rates, duplicate discordance rates can also be used as a SNP-level quality filter. The challenge lies in finding a level of discordance that would eliminate most SNPs with high error rates, while retaining most SNPs 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 7 shows these probabilities for $x = 0-4$ and $n = 204$. Here we chose $n = 204$ to correspond to the number of duplicate study sample pairs. *We recommend a filter threshold of > 2 discordant calls because this retains $> 99\%$ of SNPs with an error rate of 0.001, while removing $> 77\%$ of SNPs with an error rate 0.01.* This threshold eliminates 400 SNPs.

Figure 19 summarizes the concordance by SNP, binned by MAF. Figure 19a shows the number of SNPs in each MAF bin. Figure 19b shows the correlation (r) of allelic dosage, which is greater for SNPs with higher MAF. Figure 19c 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 19d shows the minor allele concordance, which is the concordance between genotypes in the members of sample pairs where at least one copy of the minor allele is observed (i.e. excluding pairs where both are the major homozygote). This concordance measure is more reflective of true genotyping concordance for low MAF SNPs, and the distribution is similar to the correlation of allelic dosage in that the metric increases with MAF.

13 Mendelian errors

Mendelian error rates are another way to evaluate genotype quality, both at the SNP level and at the family (trio or duo) level. For bi-allelic SNPs, Mendelian errors can only be observed between parent(s) and offspring. While many study samples are related, the relation types are full sibling, second degree, and third degree. Therefore the Mendelian error analysis was restricted to 29 HapMap trios and 2 duos. Only 0.421% of SNPs have any Mendelian errors, and only 414 SNPs have more than one error. Furthermore, all HapMap parent-offspring sets had a low burden of Mendelian errors (maximum error rate was 0.035%).

We recommend filtering out SNPs with more than one Mendelian error ($n=414$ SNPs) to avoid removing SNPs with an error in just one trio or duo, which might be due to copy number variation or other chromosomal anomaly.

14 Hardy-Weinberg equilibrium

Departure from Hardy-Weinberg equilibrium (HWE) can be used to detect SNPs with poor genotyping quality and/or artifacts. We selected a set of mutually unrelated study participants with relatively homogeneous genetic ancestry for HWE testing. The samples in this set fall within a two-dimensional ellipse of greatest density of self-reported “white” participants, based on the first two eigenvectors from the study-only PCA described in Section 9. This approach to sample selection uses the minimum covariance determinant (MCD) method [15]. The selected set consists of 6,162 unrelated subjects, which can be identified with the logical variable “pca.homog” in “Sample_analysis.csv.” We performed an additional PCA on the selected subset to check for any residual population structure. As seen in the plot of the first two eigenvectors in Figure 21, the selected sample set appears to be reasonably homogeneous and thus suitable for HWE testing.

A HWE exact test was performed at each autosomal, non-technical failure SNP, using the selected sample set described above. The subset of female samples were used to calculate HWE p-values for X chromosome SNPs. Deviations from HWE due to population structure are expected to result in an excess of homozygotes or a positive inbreeding coefficient estimate, calculated as $1 - (\text{number of observed heterozygotes}) / (\text{number of expected heterozygotes})$. A comparison of the observed distribution of the inbreeding coefficient estimates (for a random sample of autosomal SNPs) with a simulated distribution of inbreeding coefficient estimates for the same set of SNPs under the assumption of Hardy-Weinberg equilibrium was performed for 6,162 unrelated participants selected for HWE testing. Figure 22 show that the observed and simulated distributions are nearly identical in each plot. We conclude that most deviations from HWE result from genotyping artifacts, rather than population structure.

Determining a HWE threshold p-value to indicate poorly performing SNPs is somewhat subjective. The goal is to find a p-value threshold that removes the majority of SNPs with apparent genotyping artifacts while retaining the majority of SNPs without such artifacts. To arrive at a threshold, we examined many cluster plots of randomly selected SNPs within varying p-value ranges. *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 . A total of 2,672 SNPs fail this p-value threshold.*

15 Duplicate SNP probes

The OmniExpress array has four sets of SNP assays that occur as positional duplicates, as indicated by identical physical positions (chromosome and base pair) within each set. To determine whether sets of positional duplicates were genotyping consistently, we calculated concordance of genotype calls across study samples for each set. A high level of concordance indicates that these assays measure the same alleles at the same variant and thus provide redundant information, whereas low concordance suggests that the assays measure different variants and/or different alleles at the same variant site (e.g., a triallelic site where each member of a positional duplicate assays a different minor allele).

To determine a suitable cutoff for concordance, we calculated the probability of having $> x$ discordant calls over 9,014 unique study samples, given assumed error rates. We chose 149 discordances, for which the

probability is $1.11\text{e-}16$ with error rate of 0.001 and 0.989 with error rate 0.01. Pairs with ≤ 149 discordances are considered to assay the same SNP and one member of each pair (or two from triplets) is labeled as “redundant” in “SNP_analysis.csv” (the one(s) with higher MCR). Pairs with > 149 discordances are flagged as discordant by “dup.pos.disc” = TRUE in “SNP_analysis.csv.” There are two redundant SNPs and one pair of positional duplicate SNPs that are flagged as discordant.

16 Sample exclusion and filtering summary

As discussed in Section 5, genotyping was attempted for a total of 9,804 samples, of which 9,670 passed CIDR’s QC process (Table 2). The subsequent data cleaning QA/QC process identified 12 samples with questionable identity, due to irreconcilable discrepancies either between annotated and genetic sex (two samples) or relatedness: ten samples involved in expected but unobserved sibling pairs (see Section 8). An additional six samples were included in most analyses but later withdrawn from the dbGaP posting due to withdrawn consent. A further 223 samples derived from two unique WLS control non-participants are also intentionally excluded from the dbGaP posting. Therefore, 9,429 scans will be posted on dbGaP.

In general, we recommend filtering out large chromosomal anomalies associated with error-prone genotypes and whole samples with MCR $> 2\%$. In this study, all samples have MCR $\leq 2\%$, thus no samples are filtered due to high MCR. There were 35 large chromosome anomalies (31 in study samples) that are filtered, i.e. genotypes in the anomaly regions are set to missing in the filtered subject-level PLINK file provided.

We also recommend sample filters for specific types of analyses, such as PCA, HWE, and association testing, as indicated in the corresponding sections of this report. These filters generally include just one scan per participant (unduplicated) and one participant per family (unrelated) and are provided in the file “Sample_analysis.csv.”

17 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 and/or low informativeness. The suggested quality filter and composite filter are provided as logical variables in the “SNP_analysis.csv” file, which also has the individual components of these composite filters to allow users to vary thresholds as desired. The quality filters (rows 2 – 8) remove 2.27% of the 713,014 SNP assays attempted. The composite filters (rows 2 – 10), also excluding uninformative redundant and monomorphic SNPs, remove 3.42% of the SNP assays.

In addition to the composite filter, we suggest applying an allele frequency filter that also takes sample size into account (see Section 19.) For illustration, Table 1 provides figures for applying a filter of MAF < 0.01 among study subjects. The quality, informativeness, and MAF filters combined remove 12.65% of the SNP assays attempted. Note that SNP quality metrics were calculated prior to the removal of the six study samples with revoked consent.

Regardless of what filters are applied to association test results, it is highly recommended to manually review genotyping cluster plots for any SNPs of interest to ensure that the observed associations are not due to genotyping artifacts.

18 Minor allele frequency

Allele frequencies for each SNP were computed using all unique study samples. Figure 20 shows the distribution of minor allele frequency (MAF), where the minor allele is defined in all unique study samples. The percentage of all SNPs with MAF < 0.01 is 10.3% for the autosomes and 16.4% for the X chromosome.

We define “non-informative” SNPs meeting one of more of the following criteria. Items in quotes refer to variables that can be found in “SNP_analysis.csv.”

- technical failures: SNPs failed by the genotyping center, identified by “missing.n1”=1,

- redundant positional duplicates, identified by “redundant”=TRUE, or
- monomorphic in study samples, identified by “MAF.study” = 0

A total of 15,662 (2.2%) SNPs were non-informative; we refer to the remaining SNPs as informative SNPs.

Table 8 displays, by MAF bin, the total number of informative SNPs, the number of informative SNPs passing the quality filter, and the percentage of informative SNPs passing the quality filter. The quality filter is defined in Section 17. The percentages of SNPs passing the quality filter are similarly high across MAF bins. However, this does not ensure that genotype calling accuracy is equally good or better for lower MAF, because it is more difficult to identify poor performance for low MAF SNPs. For example, there is less power to detect HWE deviations and many fewer opportunities to detect Mendelian errors.

The CIDR QC process for low MAF SNPs includes running zCall [16] to identify SNPs where possible heterozygous clusters were missed by GenCall (parameters T=21 and I=0.2). SNPs with one or more possible new heterozygote points as defined by zCall and at least four total heterozygote points were manually reviewed and edited or failed as appropriate. The variable “zcall_flagged” in “SNP_annotation.csv” provides the number of new heterozygous calls recommended by the zCall algorithm.

19 Preliminary association tests

We ran preliminary association tests to assess (1) overall dataset quality, (2) adequacy of the principal components in adjusting for population structure, and (3) the effectiveness of our suggested SNP filters in removing problematic and/or low-quality SNPs. We performed linear regression on a summary cognitive measure “theta” within a maximal set of 6,543 unrelated samples. Next we describe the selection of an association model, followed by a discussion of the association test results.

For association testing, we used the same set of unique, mutually unrelated (at third degree) study samples as were used in the study-only PCA (see Section 9). Note this association analysis includes four samples later removed from the dbGaP posting. We performed linear regression to test for an association between each genotyped SNP and the outcome “theta”. Covariates were selected based on prior knowledge and by regressing potential covariates on the trait and selecting those covariates that had a significant relationship with the outcome. The results of regressing sex, age, and the first 12 EVs are shown in Table 9. Here age is the age calculated at a consistent time point (versus sample age), where extreme ages have been winsorized to protect participant privacy. We decided to include EV1-6, as no higher EVs were significant after a multiple testing correction. The final model is:

$$theta \sim sex + age + EV1 + EV2 + EV3 + EV4 + EV5 + EV6 + SNP\ genotype.$$

For autosomal SNPs, all sample genotypes were coded as 0, 1, or 2 copies of the A allele, where the A allele was defined via Illumina allele nomenclature [17]. In performing association tests for X chromosome 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. The outcome “theta” is a normalized measure that ranges from -2 to 2.

Figure 23 shows QQ plots for likelihood ratio tests of SNP effect on the outcome. Results are given with no SNP filter, with the recommended composite (quality plus informativeness) filter, and with the composite filter plus an “effective sample size filter” of $2p(1-p)N > 30$, where p is the minor allele frequency and N is the number of samples. This yields a MAF filter of $MAF > 0.002$. With the recommended filtering (lower left plot), the QQ plot shows no appreciable inflation ($\lambda = 1.02$). Note the highest λ is for low MAF SNPs (lower right plot), further supporting the use of a MAF filter.

The corresponding Manhattan plot is shown in Figure 24. Similarly to the QQ plots, Manhattan plots are shown with no SNP filter (top row), with the composite filter (middle row), and with the composite plus effective sample size filter (bottom row). These plots show that there is not an excess of positive hits and those loci that show evidence of association (albeit not meeting the genome-wide significance threshold)

have evidence from multiple SNPs. These characteristics suggest that there are not artifacutal or spurious association signals.

As an additional QC measure, we examined cluster plots of the most significant SNPs (see Figure 25). These cluster plots are of high quality, suggesting that the statistical association is not driven by genotyping artifacts. Thus we consider the results of these preliminary association tests to indicate a high quality dataset and effective SNP quality filters. Results for these association tests are provided in the file “assoc_results.csv.”

Appendix

A Project participants

Department of Sociology, University of Wisconsin - Madison

Pamela Herd, Carol Roan, Kamil Sicinski, and Huey-Chi Vicky Chang

Northwestern University

Yoonjung Joo, Fred Lin, and M. Geoffrey Hayes

Center for Inherited Disease Research, Johns Hopkins University

Kim Doheny, Jane Romm, Hua Ling, and Elizabeth Pugh

Genetic Analysis Center, Department of Biostatistics, University of Washington

Sarah Nelson, Cecelia Laurie, Cathy Laurie, Bruce Weir

dbGaP-NCBI, National Institutes of Health

Nataliya Sharopova

References

- [1] S. M. Gogarten, T. Bhangale, M. P. Conomos, C. A. Laurie, C. P. McHugh, I. Painter, X. Zheng, D. R. Crosslin, D. Levine, T. Lumley, S. C. Nelson, K. Rice, J. Shen, R. Swarnkar, B. S. Weir, and C. C. Laurie. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics*, 28(24):3329–3331, Dec 2012.
- [2] Matthew P. Conomos and Timothy Thornton. *GENESIS: GENetic ESTimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness*, 2016. R package version 2.1.7.
- [3] X. Zheng, D. Levine, J. Shen, S. M. Gogarten, C. Laurie, and B. S. Weir. A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics*, 28(24):3326–3328, Dec 2012.
- [4] C.C. Laurie et al. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34:591–602, 2010.
- [5] D.A. Peiffer et al. High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research*, 16:1136–1148, 2006.
- [6] 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.
- [7] 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.

- [8] 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.
- [9] Ani Manichaikul, Josyf C. Mychaleckyj, Stephen S. Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.
- [10] N. Patterson, A.L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS Genetics*, 2:e190, 2006.
- [11] J. Novembre et al. Genes mirror geography within Europe. *Nature*, 456:98–101, 2008.
- [12] K. A. Frazer et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*, 449(7164):851–861, Oct 2007.
- [13] Xiaofeng Zhu, S. Li, R. S. Cooper, and R. C. Elston. A unified association analysis approach for family and unrelated samples correcting for stratification. *American Journal of Human Genetics*, 82:352–365, 2008.
- [14] 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.
- [15] P.J. Rousseeuw and K. van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.
- [16] J. I. Goldstein, A. Crenshaw, J. Carey, G. B. Grant, J. Maguire, M. Fromer, C. O’Dushlaine, J. L. Moran, K. Chambert, C. Stevens, P. Sklar, C. M. Hultman, S. Purcell, S. A. McCarroll, P. F. Sullivan, M. J. Daly, and B. M. Neale. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. *Bioinformatics*, 28(19):2543–2545, Oct 2012.
- [17] S. C. Nelson, K. F. Doheny, C. C. Laurie, and D. B. Mirel. Is ‘forward’ the same as ‘plus’?...and other adventures in SNP allele nomenclature. *Trends Genet.*, 28(8):361–363, Aug 2012.

Table 1: Summary of recommended SNP filters. The number of SNPs lost is given for sequential application of the filters in the order given. For a description of the criteria for CIDR technical failures, refer to the CIDR document “SNP_Summary_README.pdf.” Rows 2 - 10 comprise the composite.filter, which is a combination of quality metrics (rows 2 - 8) and informativeness (rows 9 - 10). The sex difference metrics in lines 7 and 8 were computed on the homogeneous genetic ancestry sample set identified by “pca.homog” in “Sample_analysis.csv” (see Section 14).

Filter	SNPs.lost	SNPs.kept
1 none - all SNP probes		713,014
2 CIDR technical filters	7,418	705,596
3 Missing call rate $\geq 2\%$	6,181	699,415
4 > 2 discordant calls in 204 study duplicates	85	699,330
5 > 1 Mendelian error in 31 trio/duo sets	405	698,925
6 HWE p-value $< 10^{-4}$ in homogeneous sample set	2,063	696,862
7 Sex difference in allele freq ≥ 0.2 for autosomes/XY	42	696,820
8 Sex difference in heterozygosity > 0.3 for autosomes/XY	0	696,820
9 positional duplicates	2	696,818
10 MAF = 0 across all study samples	8,222	688,596
11 MAF < 0.01	65,769	622,827
12 Percent of SNPs lost due to quality filter (rows 2-8)	2.27%	
13 Percent of SNPs lost due to composite filter (rows 2-10)	3.42%	
14 Percent of SNPs lost due to composite filter and MAF (rows 2-11)	12.65%	

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

	Study	HapMap	Both
DNA samples into genotyping production	9,606	198	9,804
Failed samples	-134	0	-134
Scans released by genotyping center	9,472	198	9,670
Scans failing post-release QC	0	0	0
Scans with unresolved identity issues	-12	0	-12
Scans with revoked consent	-6	0	-6
WLS control non-participant	-223	0	-223
Scans to post on dbGaP	9,231	198	9,429

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

	Study	HapMap	Both
Scans to post on dbGaP	9,231	198	9,429
Subjects	9,012	97	9,109
Replicated subjects	219	97	316
Families (N > 1)	2,263	27	2,290
Singletons	4,601	12	4,613

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

k_2	k_1	k_0	Kinship	Relationship
1.00	0.00	0.00	0.5	monozygotic 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: p values for regression of the summary cognitive score “theta” on each of the first twelve eigenvectors sequentially. Regression analysis included 6,543 subjects, selected as described in Section 19. The values in column “significance” indicates the level of significance where * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$

Eigenvector	p-value	significance
EV1	0.18	
EV2	1.9e-07	**
EV3	0.62	
EV4	0.027	*
EV5	0.41	
EV6	0.0001	**
EV7	0.1	
EV8	0.068	.
EV9	0.18	
EV10	0.45	
EV11	0.85	
EV12	0.51	

Table 6: Summary of SNP genotyping failures and missingness by chromosome type. A=autosomes, X=chromosome X, XY=pseudoautosomal, Y=chromosome Y, U=unknown position. The row “SNP technical failures” gives the percentage 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	X	XY	Y	U
number of probes	693,051	17,298	887	1,136	642
SNP tech failures	0.008242	0.061105	0.258174	0.206866	0.288162
missing > 0.05	0.001276	0.000062	0.000000	0.000000	0.024070

Table 7: Probability of observing more than the given number of discordant calls in 204 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 is > 2 discordant calls.

# discordant calls	error=1e-05	error=1e-04	error=1e-03	error=1e-02	# SNPs
> 0	4.07e-03	4.00e-02	3.35e-01	0.98	9,598
> 1	8.26e-06	8.06e-04	6.35e-02	0.91	1,158
> 2	1.11e-08	1.08e-05	8.26e-03	0.77	400
> 3	1.12e-11	1.09e-07	8.13e-04	0.58	148

Table 8: Summary of number and quality of SNPs by MAF bin for informative SNPs, i.e. after removing failed, redundant, and monomorphic SNPs.

	(0,0.01]	(0.01,0.05]	(0.05,0.5]	All
# of SNPs	66,035	46,323	584,994	697,352
# passing quality.filter	65,769	45,919	576,908	688,596
% passing quality.filter	99.6%	99.13%	98.62%	98.74%

Table 9: Regression of the cognitive summary score “theta” on potential covariates for 6,543 samples. Eigenvectors EV1 – EV12 are eigenvectors from the study-only PCA. “pval” is the p-value from the regression; signif indicates the level of significance where * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$.

covar	pval	signif
sex	0.02	*
age.cons	2.8e-21	****
EV1	0.15	
EV2	7.9e-08	****
EV3	0.52	
EV4	0.022	*
EV5	0.35	
EV6	8.3e-05	****
EV7	0.1	
EV8	0.063	
EV9	0.13	
EV10	0.45	
EV11	0.74	
EV12	0.58	

Figure 1: These scatterplots illustrate a check for consistency between annotated and genetic sex. Each point is a study sample, color-coded according to annotated (i.e., expected) sex. 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 the give chromosome. Sample sizes (number of SNP probes) are reported 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.

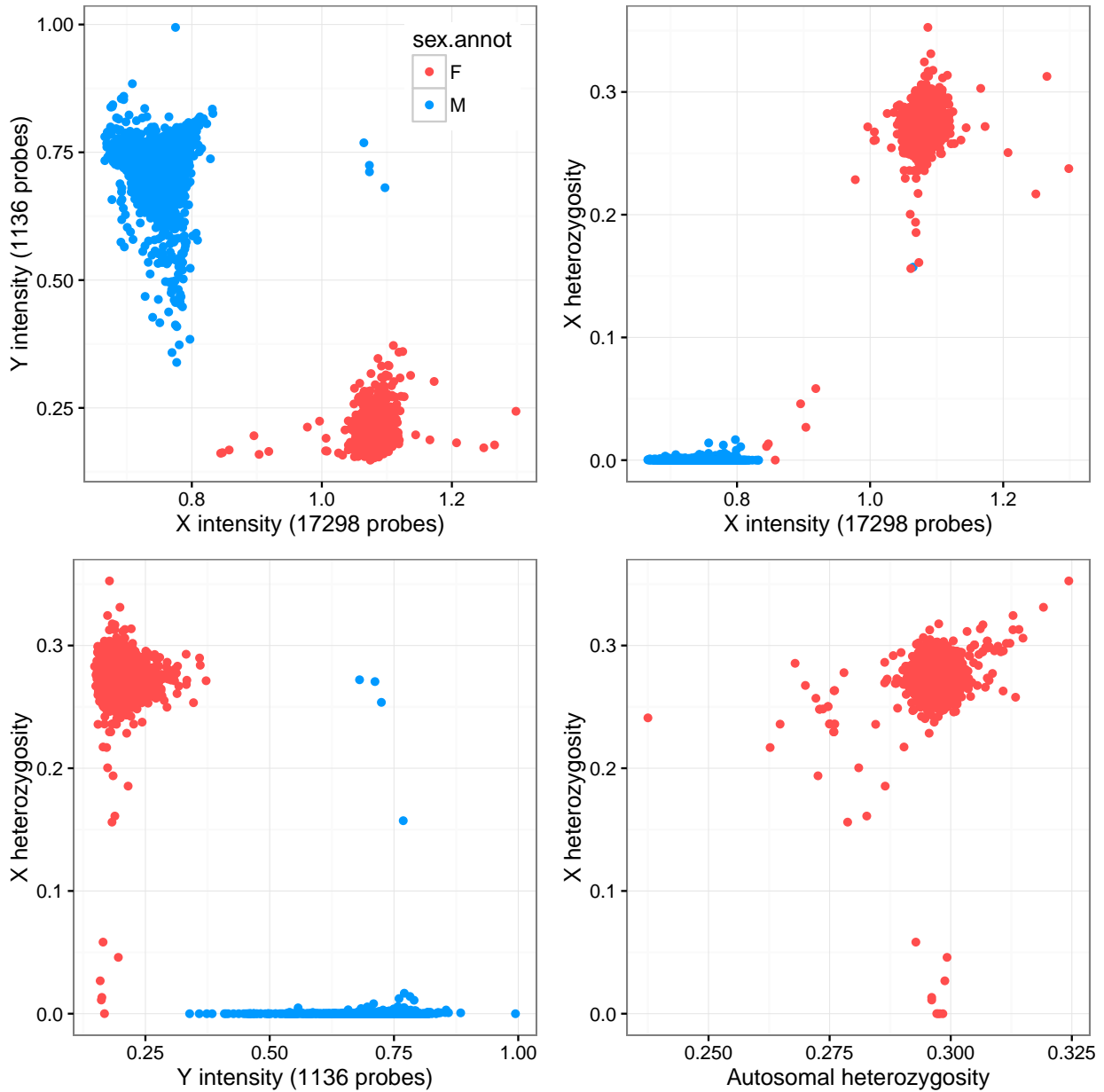


Figure 2: Potential sex chromosome anomalies. The X and Y intensities are calculated for each sample as the mean of the sum of the normalized intensities of the two alleles for each probe on the given chromosome. Groups of samples with possible sex chromosome anomalies are color-coded by the corresponding likely karyotype.

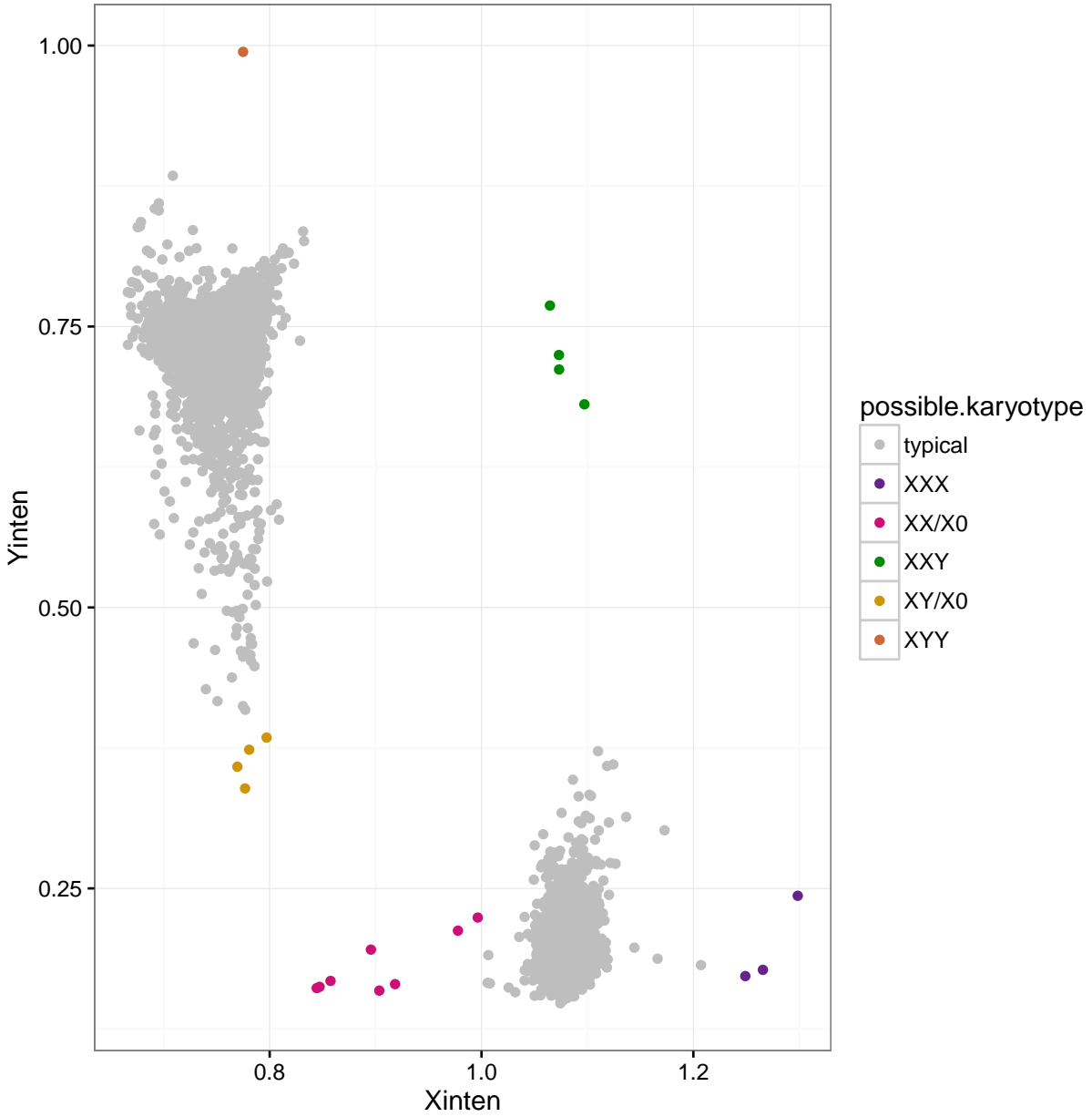


Figure 3: LRR and BAF plots for chromosome 10 in Sample A. This chromosome shows a typical pattern. In the BAF plot (lower panel), color-coding indicates genotype call: orange=AA, green=AB, fuschia=BB, black=missing. The red box on the chromosome ideogram indicates the region shown in the BAF and LRR plots.

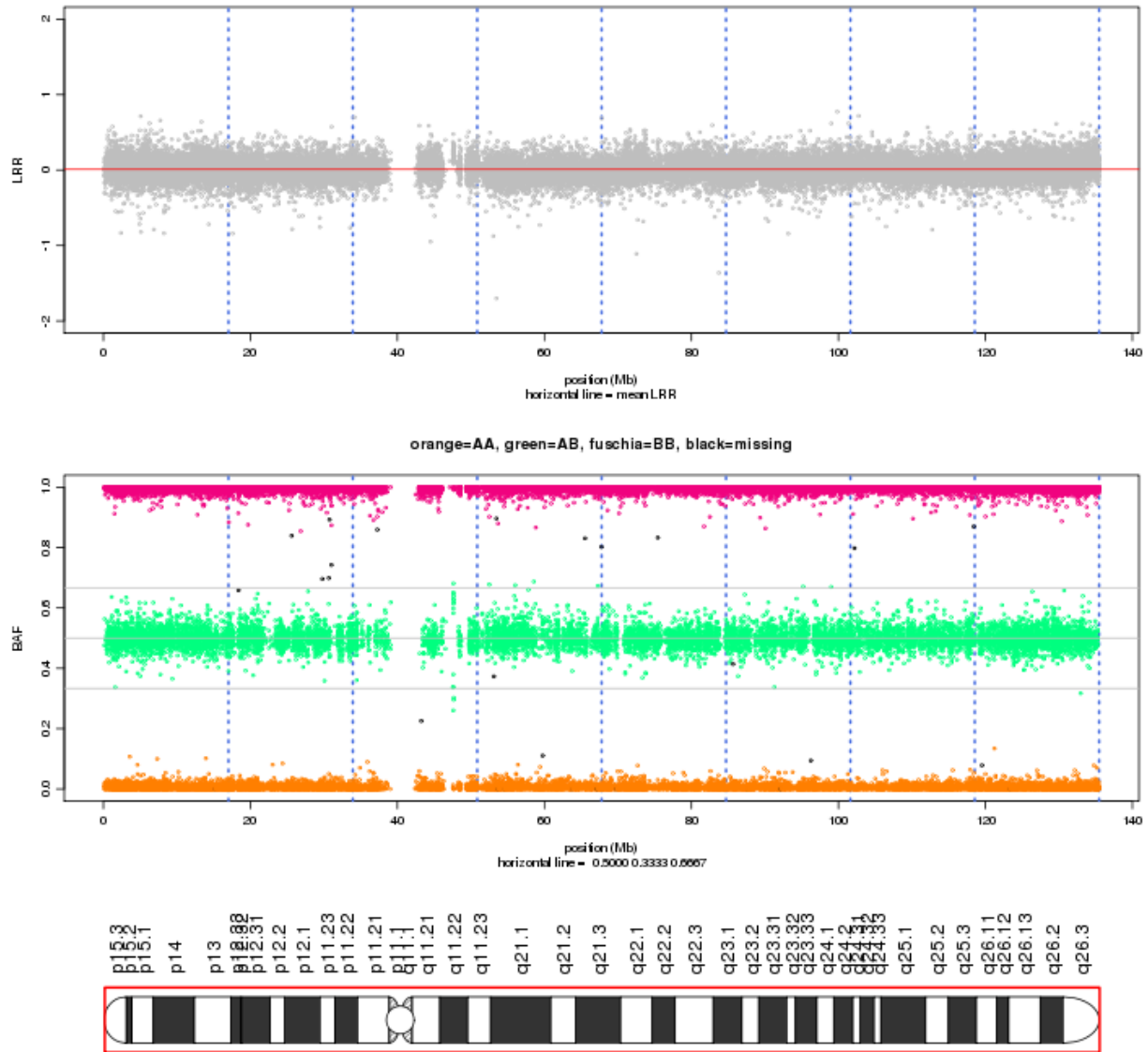


Figure 4: LRR and BAF plots for chromosome 9 in Sample A. This chromosome shows a BAF and LRR pattern consistent with a partial mosaic UPD. LRR is unchanged, though there is a split in the heterozygous band wide enough to potentially cause genotyping errors. In the BAF plot (lower panel), color-coding indicates genotype call: orange=AA, green=AB, fuschia=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. The red box on the chromosome ideogram indicates the region shown in the BAF and LRR plots.

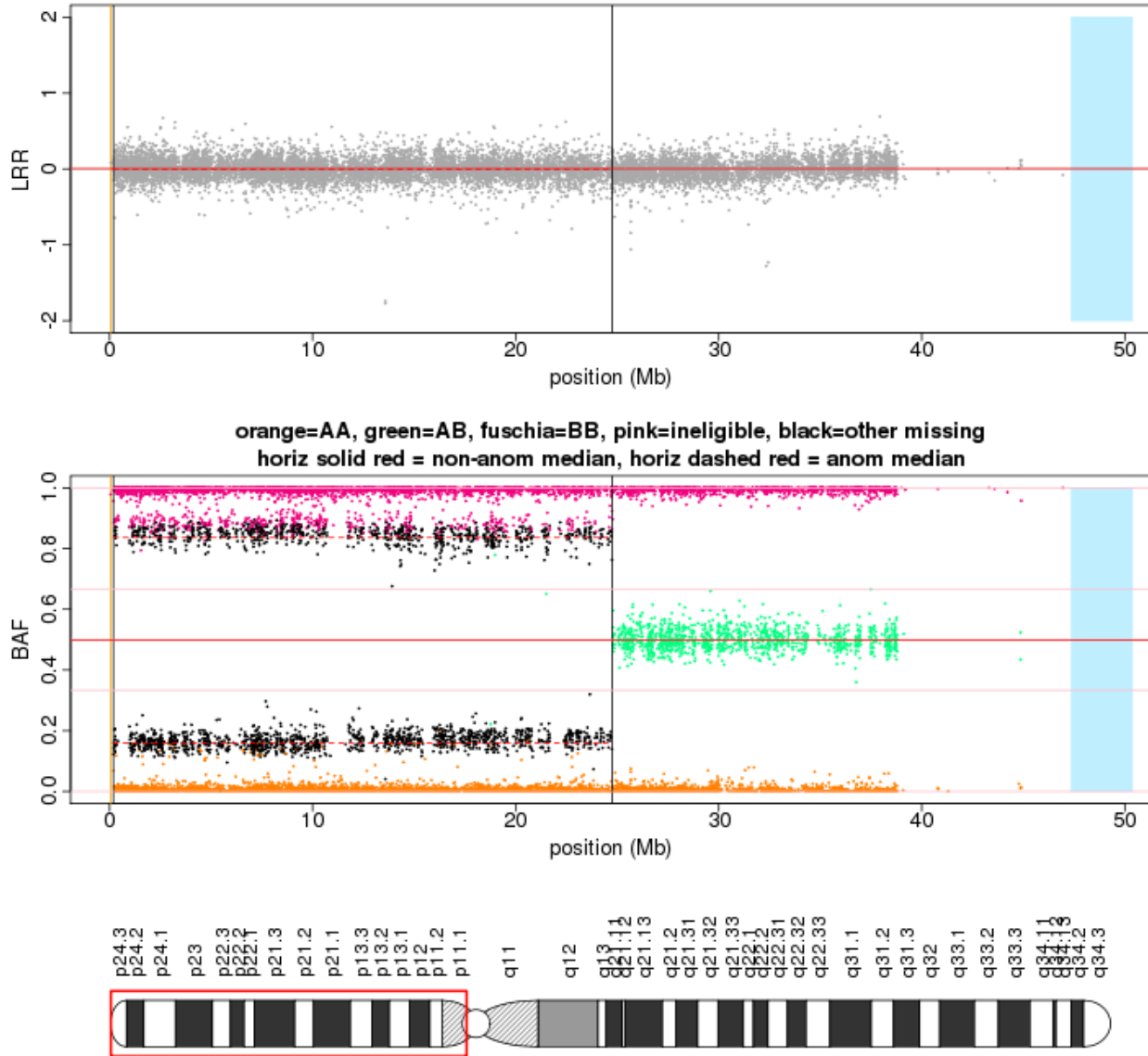


Figure 5: LRR and BAF plots for chromosome X in an XX/X0 female sample recommended for filtering. The pattern of the BAF/LRR plots is consistent with an XX/X0 karyotype, where a subset of cells are missing an entire X chromosome. LRR is decreased across the entire length of the chromosome, and the split in the heterozygous band seen in the BAF plot is wide enough to result in some truly heterozygous genotypes miscalled as homozygous.

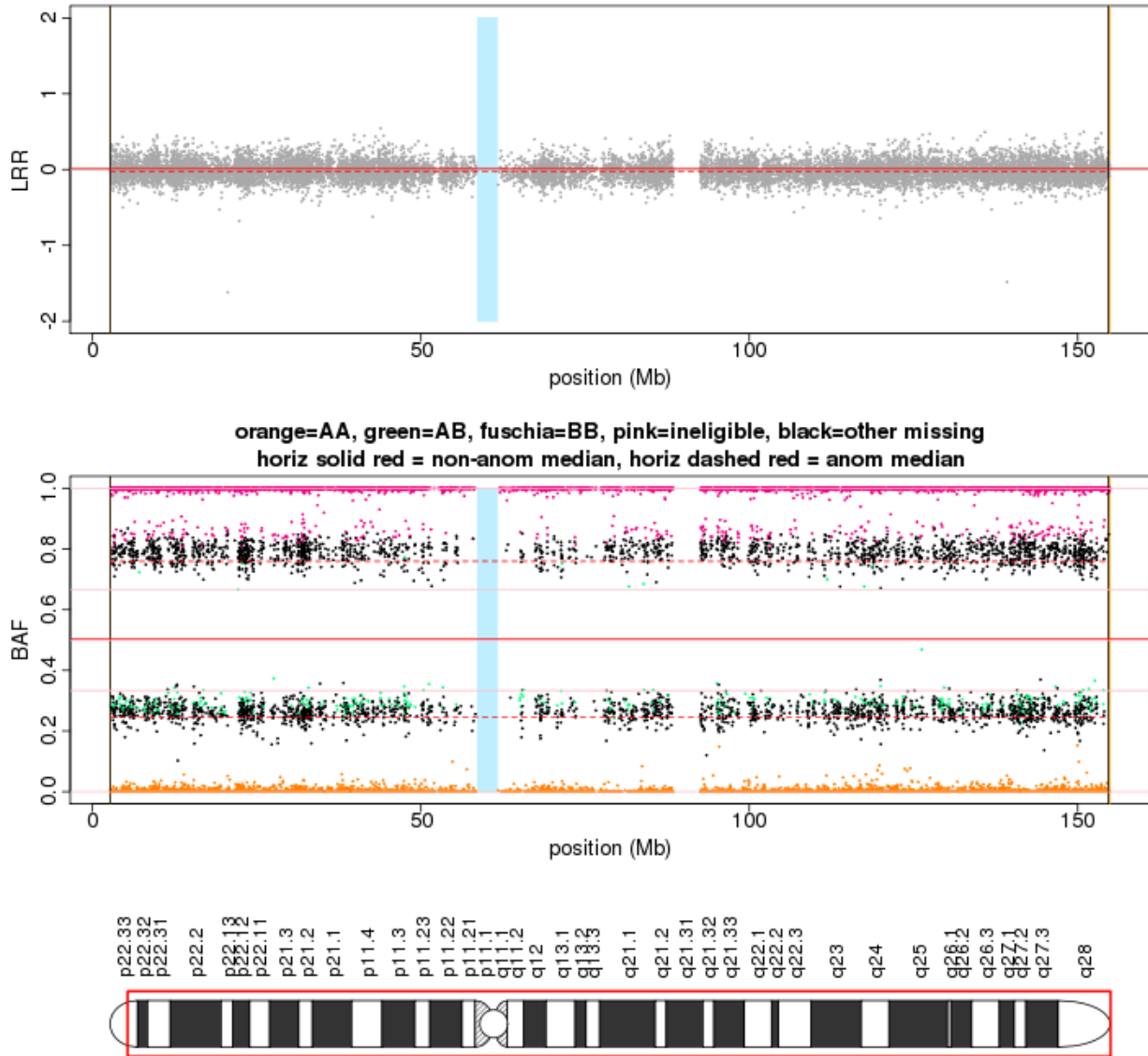


Figure 6: LRR and BAF plots for chromosome X in an XX/X0 female sample not recommended for filtering. The relatively narrow split in the BAF band indicates a high ratio of XX:X0 cells in the sample, which does not lead to erroneous genotype calls.

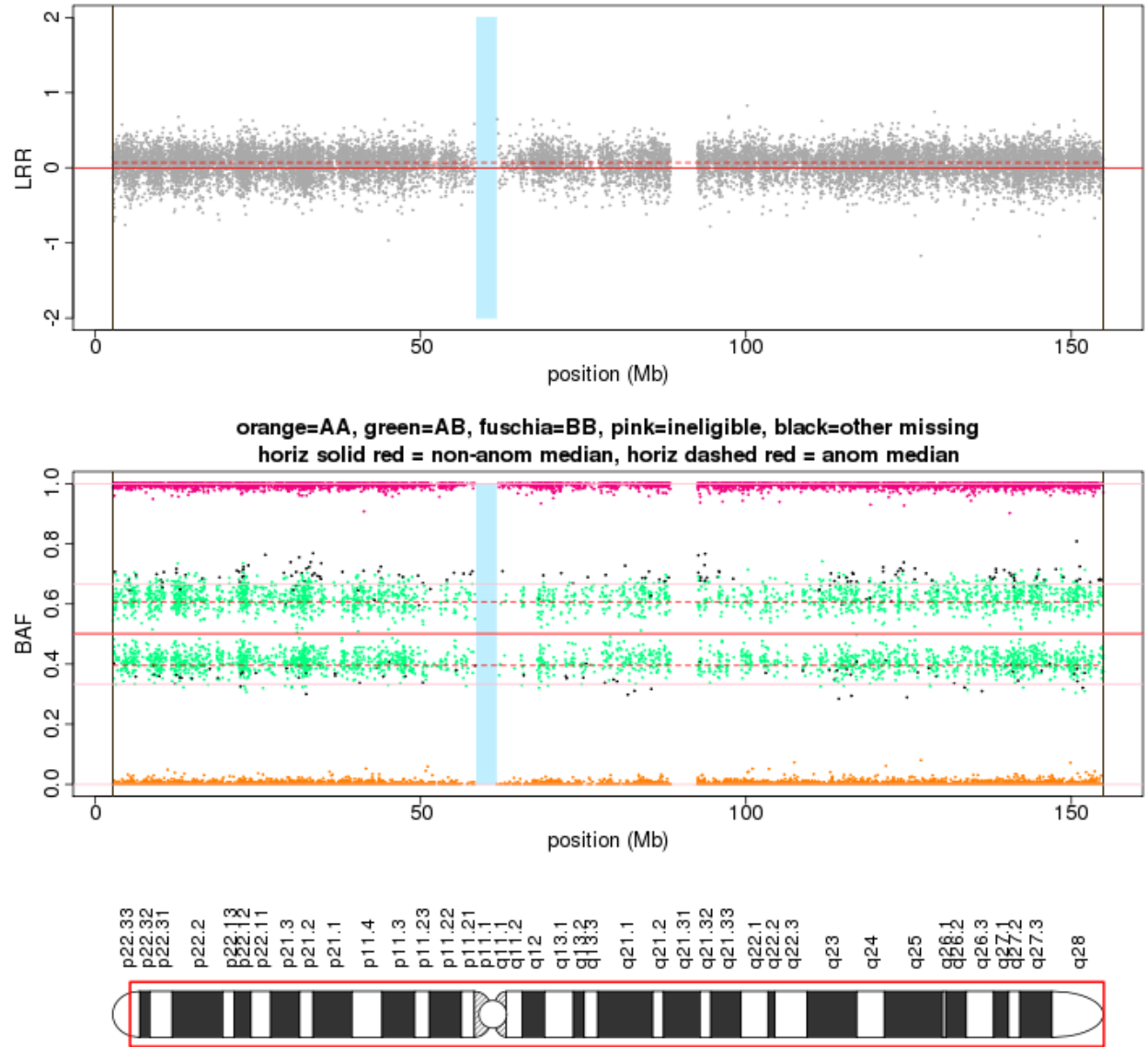


Figure 7: LRR and BAF plots for chromosome X in a male sample with a likely XXY karyotype. The pattern of the BAF/LRR plots is consistent with an XXY karyotype (see narrative for details). Color-coding is green for SNPs in the pseudo-autosomal regions (PAR1 and PAR2, shown in gray rectangles, and XTR, shown in a yellow rectangle) and fuschia for other X chromosome SNPs.

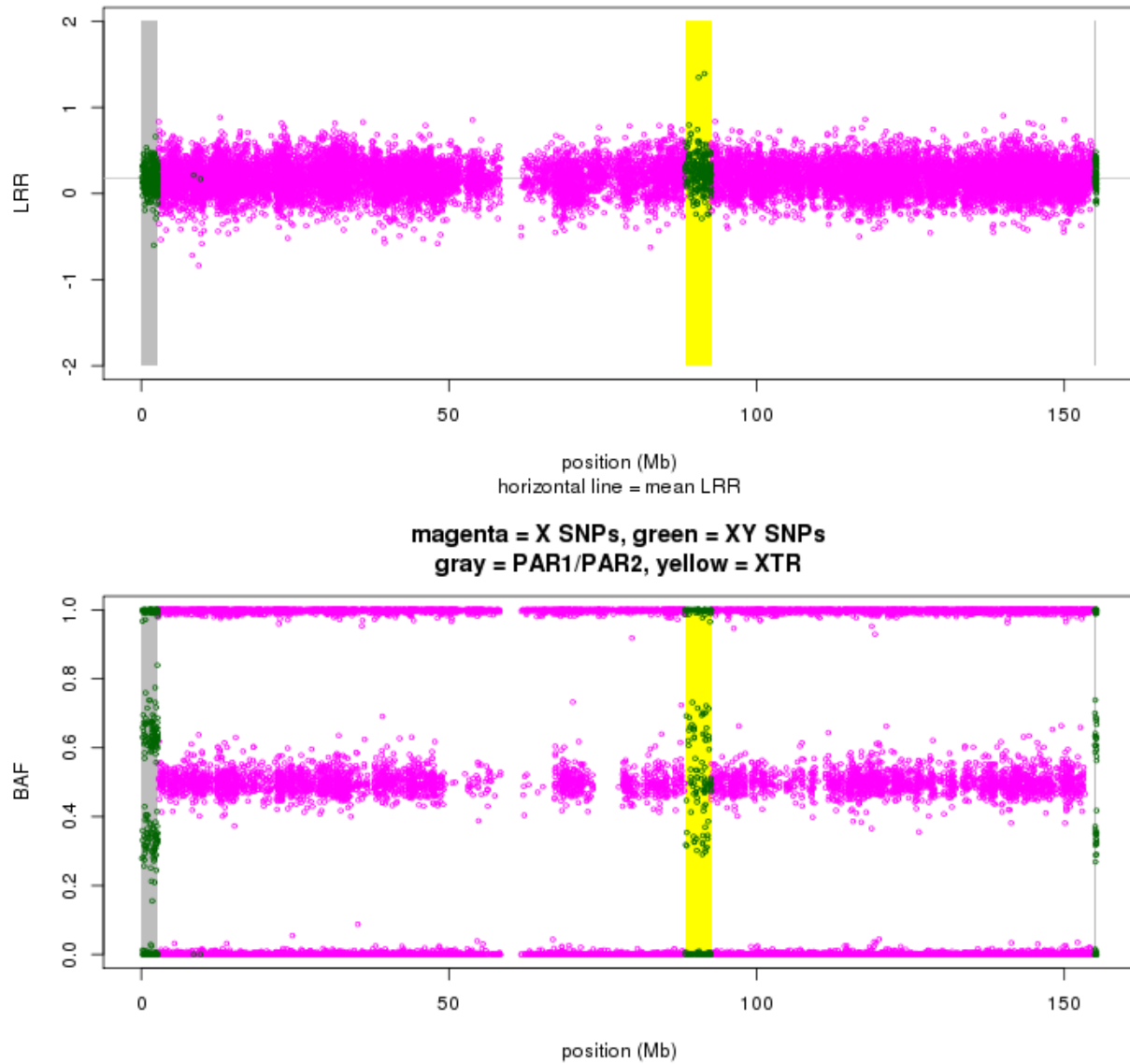


Figure 8: IBD coefficients to estimate relatedness. This plot shows 15,917 pairs of study participants and HapMap controls with an estimated $KC > 1/32$. Each point represents a pair of samples. Symbol shape denotes expected (circles) versus unexpected (triangle) relationships, and symbol color denotes expected relationship type, prior to pedigree resolutions (see report narrative for detail). Gray dashed horizontal lines show boundaries for KC values for inferring varying degrees of relatedness. Moving from the top down, the first and second dashed lines form a region for expected full siblings, the second and third form a region for expected second-degree relatives, the third and fourth for expected third-degree relatives and below the fourth we expect unrelated or related at fourth or higher degree. (See Table 1 in [9].) The vertical dashed gray line represents a tolerance for designating PO pairs or duplicates, whose IBS0 is theoretically 0. In the legend, “Dup” = duplicates, “PO” = parent-offspring, “FS” = full siblings, “Deg2” = second degree (half-sibling/avuncular/grandparent-grandchild), and “Deg3” = third degree (e.g., first cousins), and “Unrel” = unrelated samples.

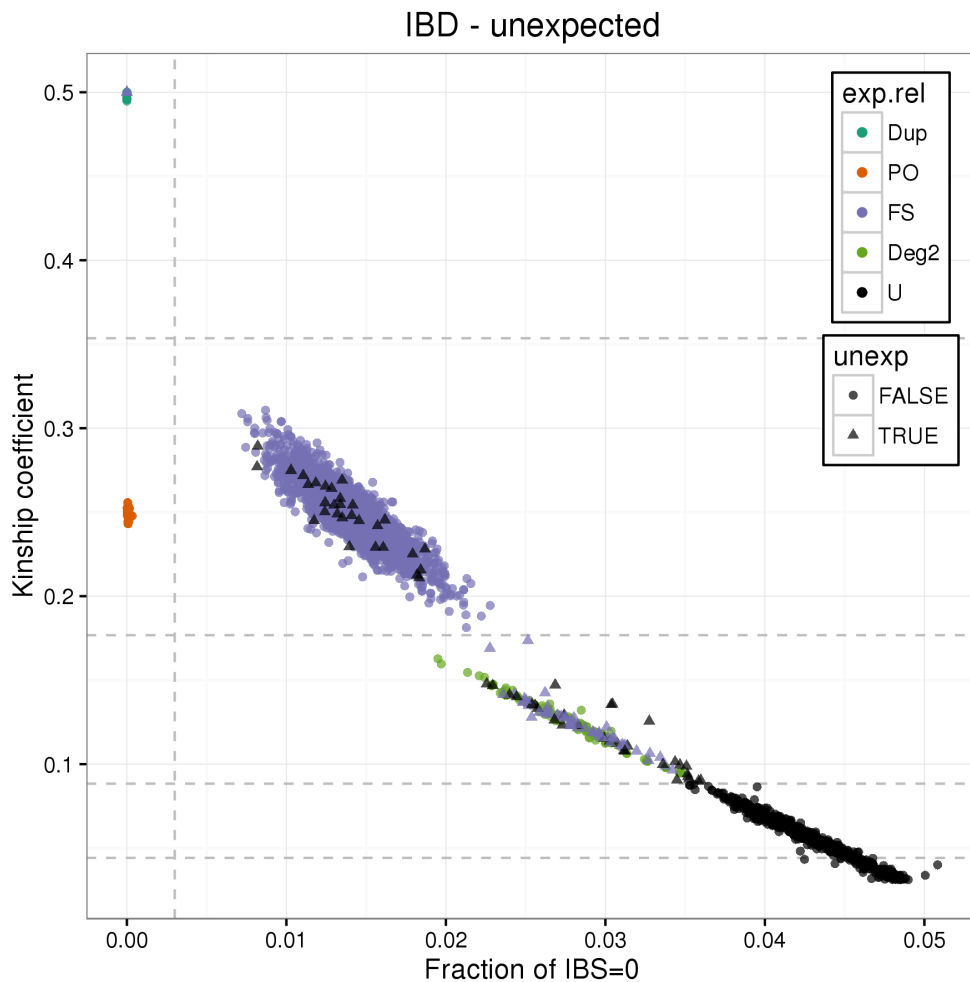
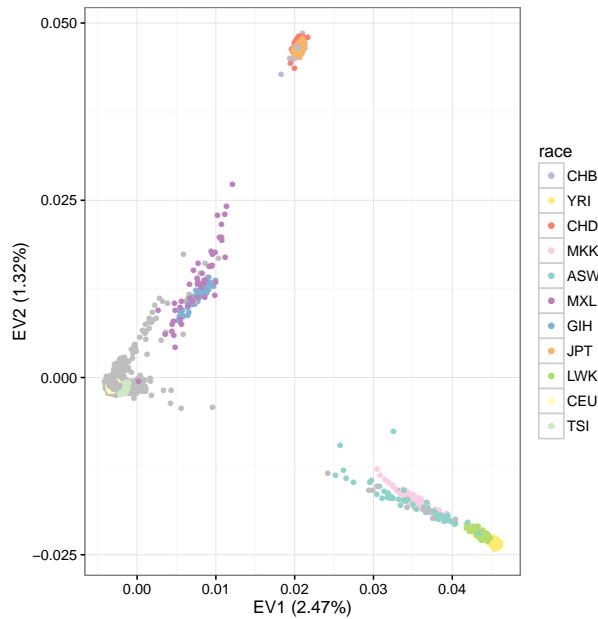
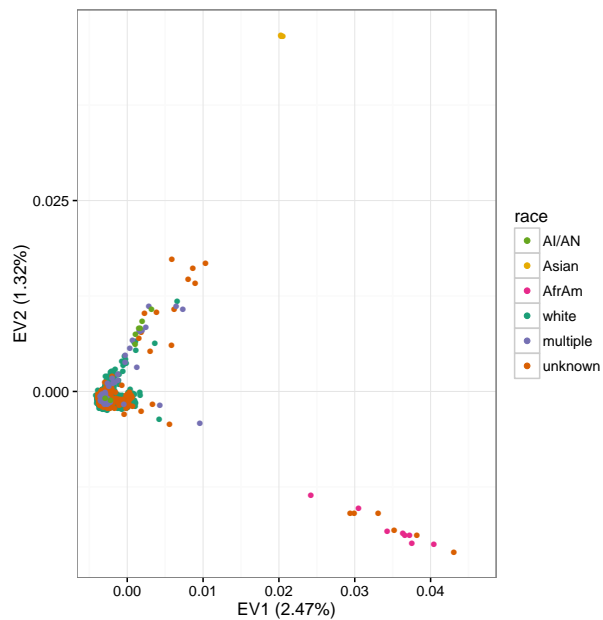


Figure 9: Principal component analysis of 9,018 study participants with 1,138 externally genotyped population controls from HapMap 3. Separate plots (on the same scale) of HapMap controls and study subjects are provided for ease of visual comparison. In the top plot, HapMap samples are color-coded according to population (see Section 9 for population descriptors); study samples are plotted in light gray. The bottom plot is limited to study samples only, color-coded by self-identified race group (AI/AN=American Indian/Alaska Native; AfrAm=Black or African American; multiple=More Than One Race; unknown=Unknown or Not Reported). Axis labels indicate the percentage of variance explained by each eigenvector.



(a) HapMap 3 controls



(b) Study participants

Figure 10: Principal component analysis of 6,543 unrelated study samples without HapMap controls. Samples are color-coded by self-identified race group (AI/AN=American Indian/Alaska Native; AfrAm=Black or African American; multiple=More Than One Race; unknown=Unknown or Not Reported). Axis labels indicate the percentage of variance explained by each eigenvector.

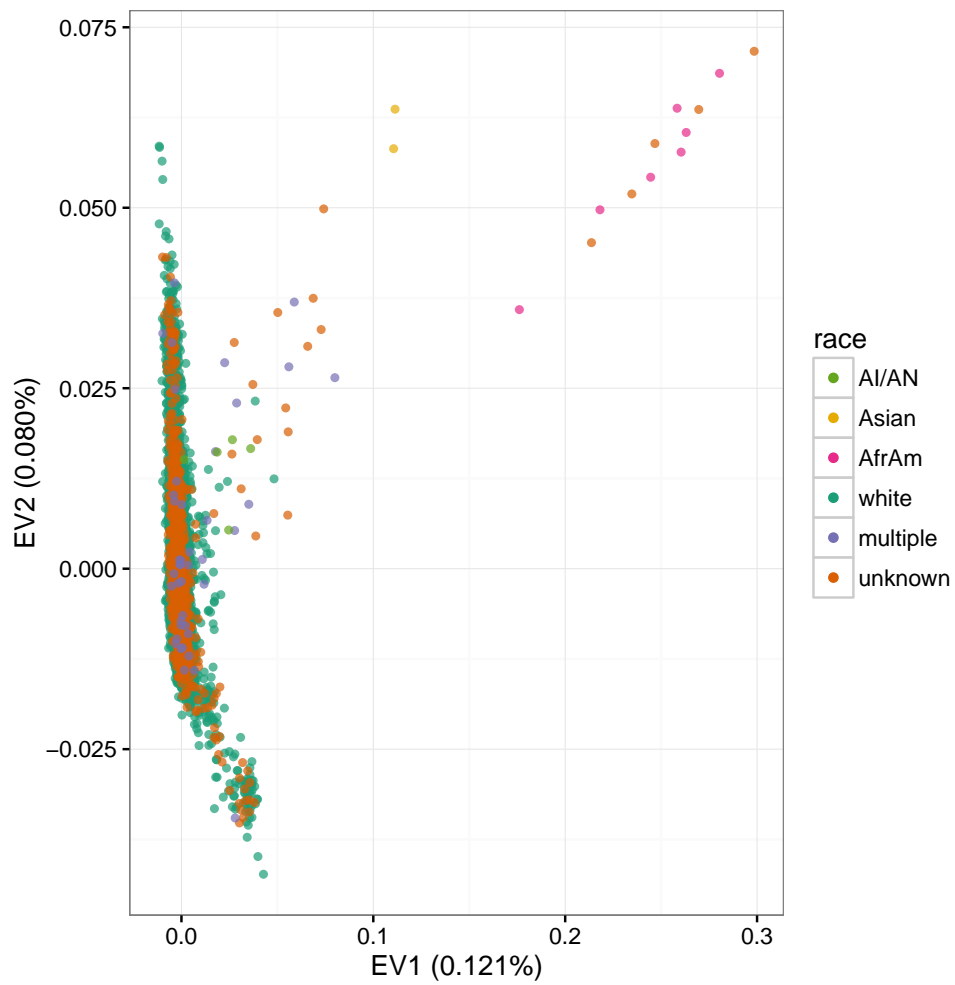


Figure 11: SNP position versus correlation between SNP genotype (0, 1 or 2 copies of the A allele) and each of the first 8 eigenvectors. These eigenvectors are from the PCA of all unrelated study participants.

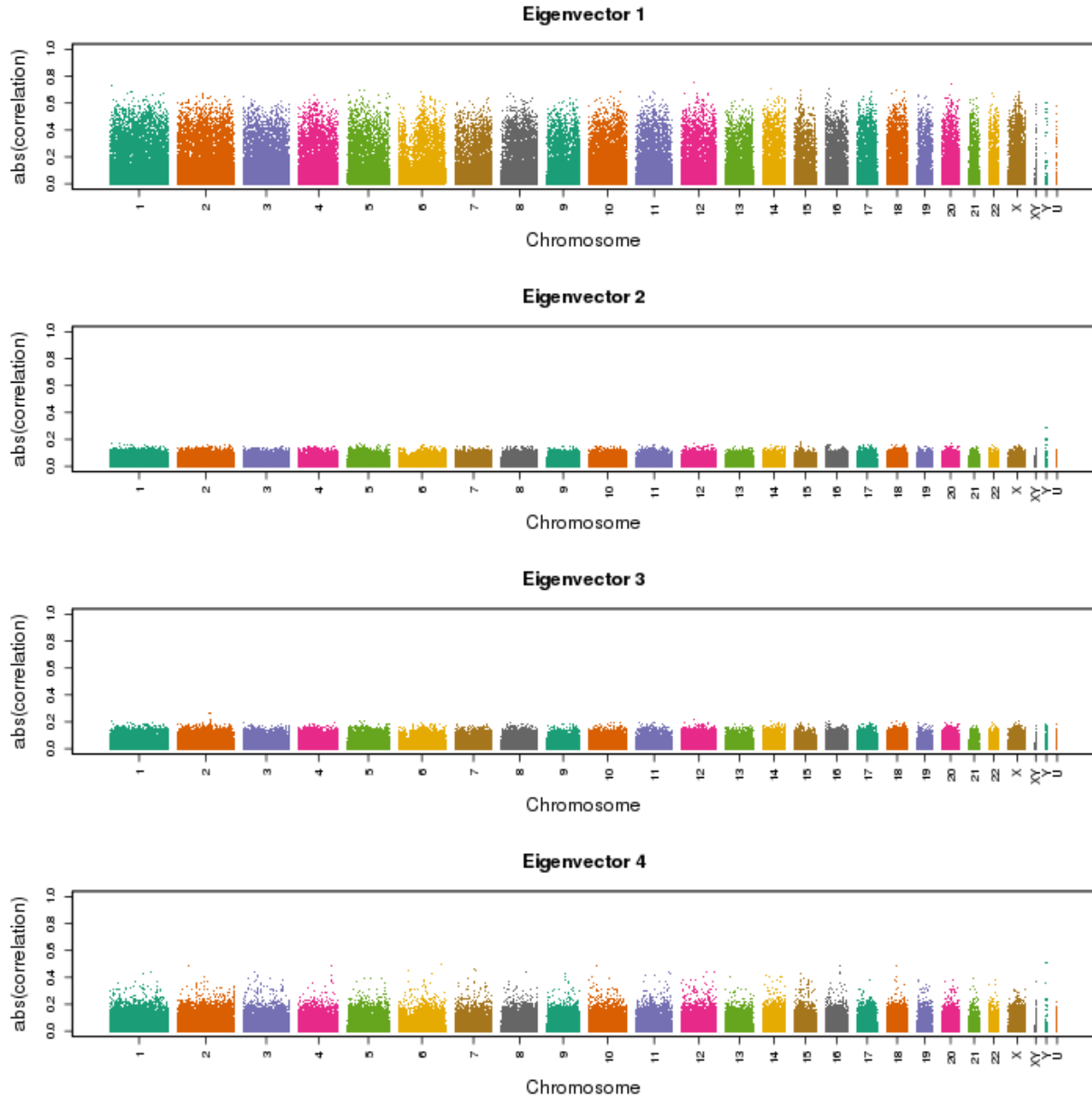


Figure 11: Continued.

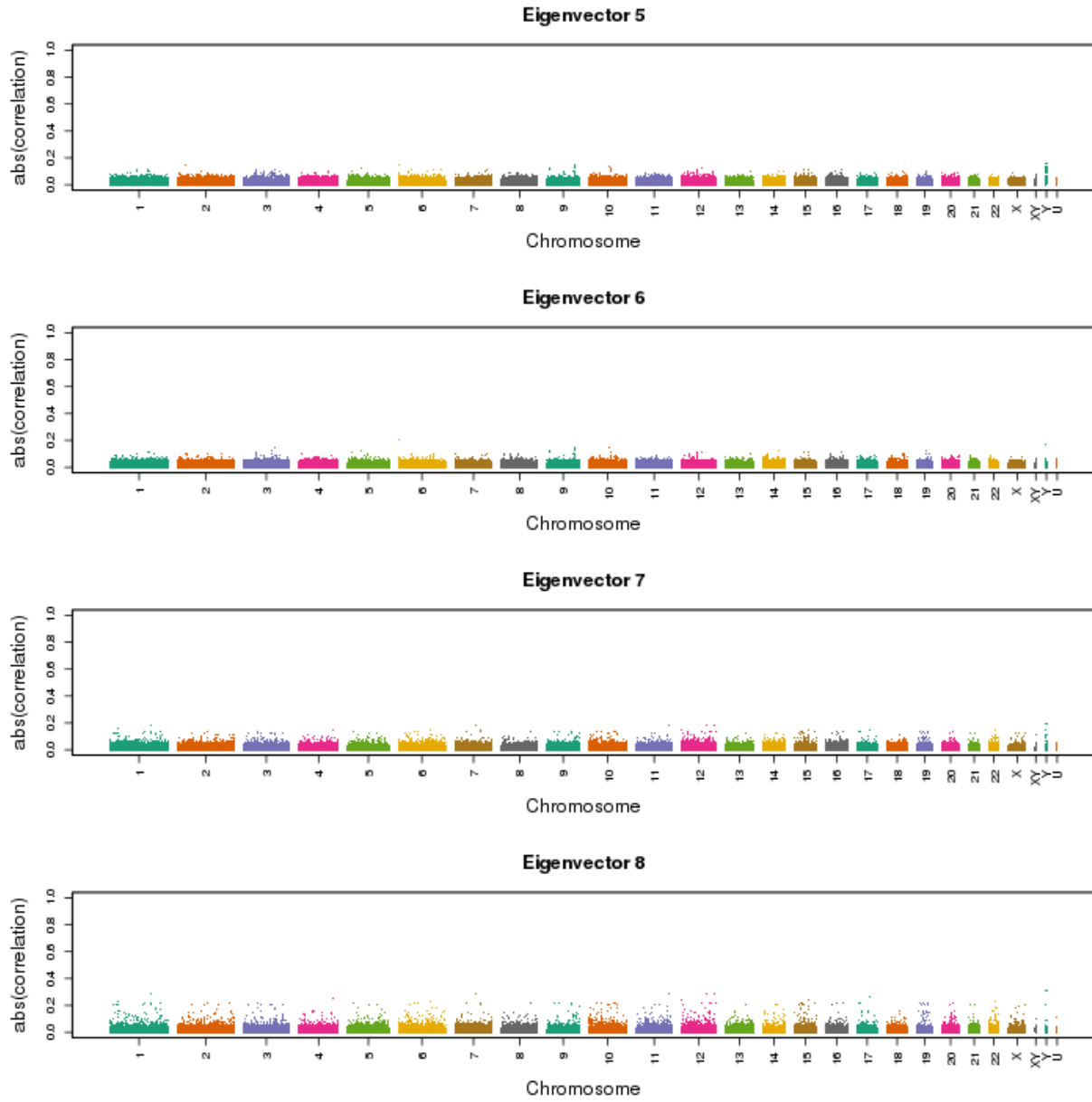


Figure 12: Scree plot for PCA shown in Figure 10.

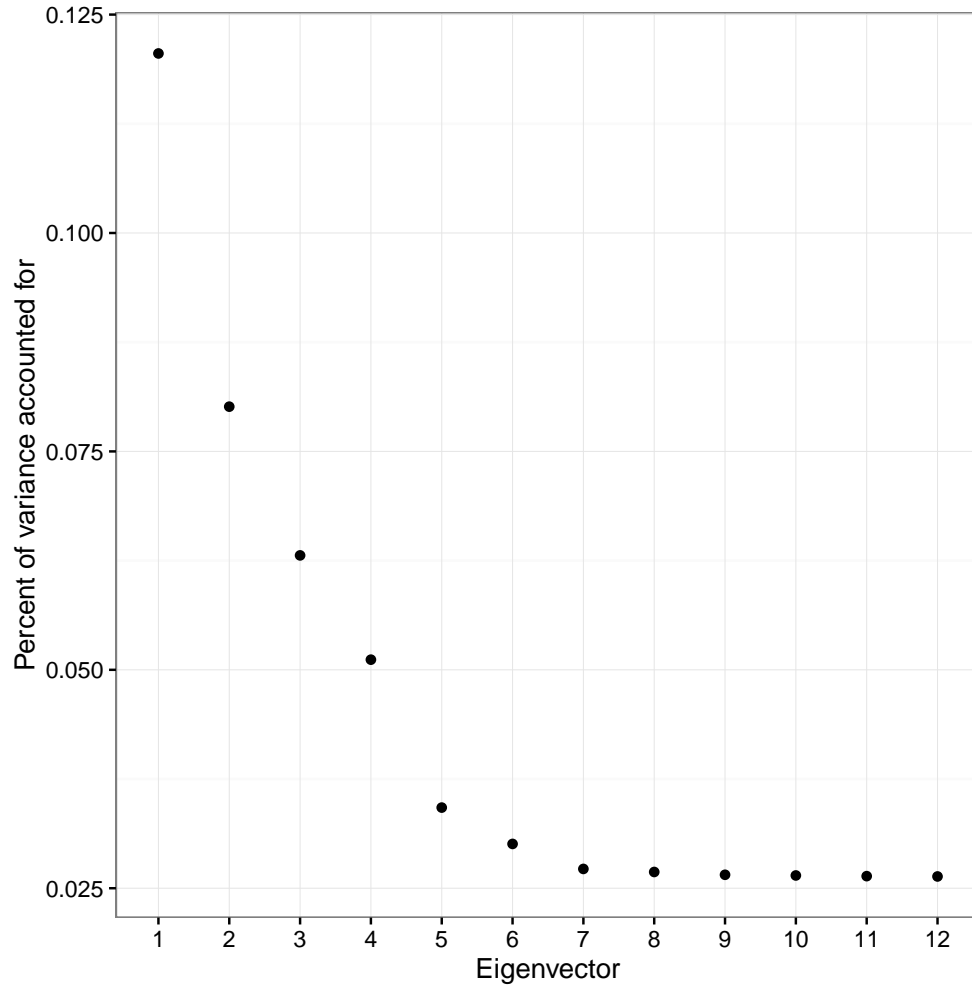


Figure 13: Parallel coordinates plot for visualizing the relationship between self-identified race and the first 12 eigenvectors from the study only PCA (see Figure 10). Vertical lines represent eigenvectors and each piece-wise line between the vertical lines traces eigenvector values for a given subject. Color-coding is according to the self-reported race (AI/AN=American Indian/Alaska Native; AfrAm=Black or African American; multiple=More Than One Race; unknown=Unknown or Not Reported).

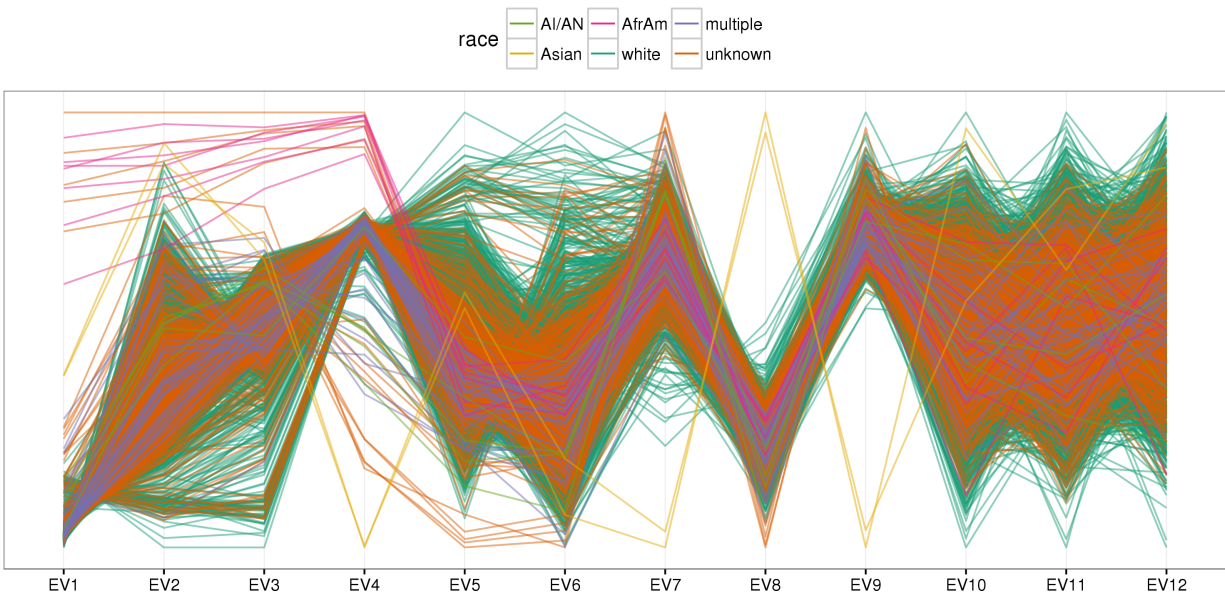


Figure 14: Principal component analysis of 6,543 unrelated study samples (left panel) along with 2,475 relatives (right panel). Samples are color-coded according to self-identified race (AI/AN=American Indian/Alaska Native; AfrAm=Black or African American; multiple=More Than One Race; unknown=Unknown or Not Reported).

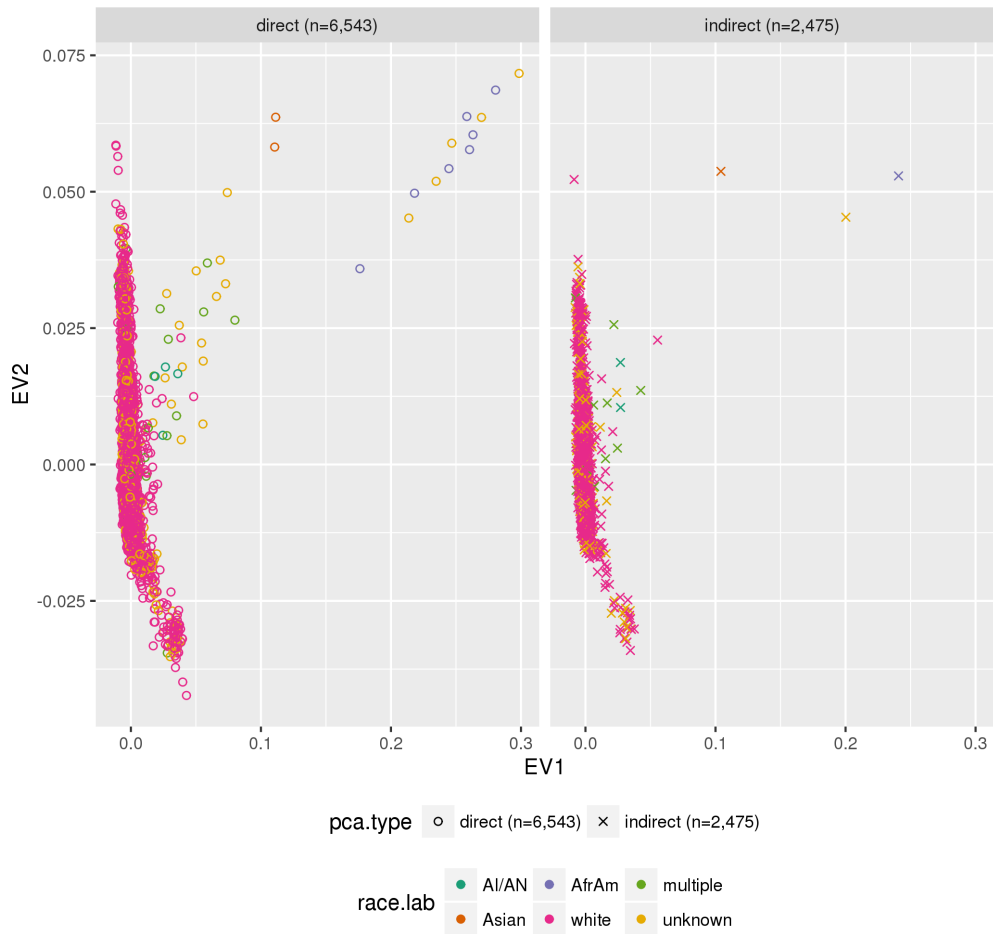


Figure 15: Histogram of the missing call rate per sample (*missing.e1*).

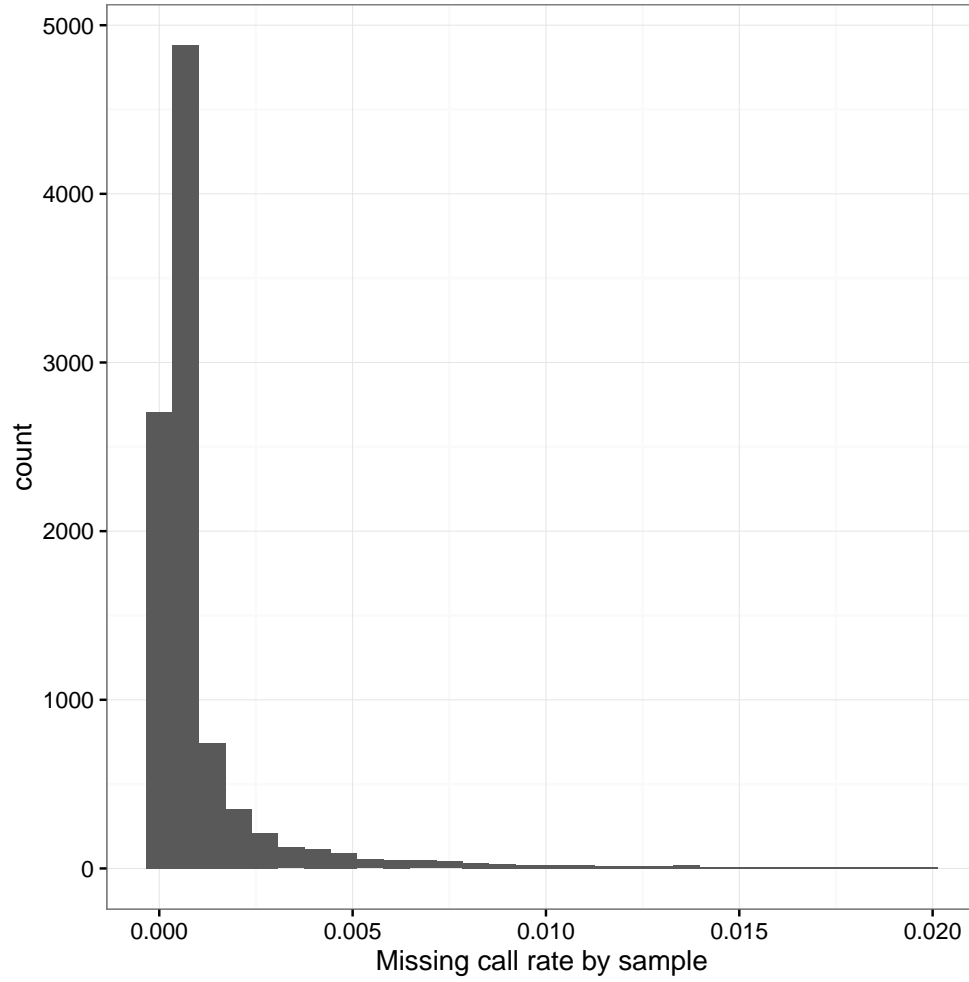


Figure 16: Histogram of the missing call rate by sample collection batch. Batch 1 samples are from a 2008-2009 mail collection; batch 2 samples were collected during in-person interviews from 2011-2013. For both batches, DNA was extracted from stored saliva using a modified Oragene extraction protocol.

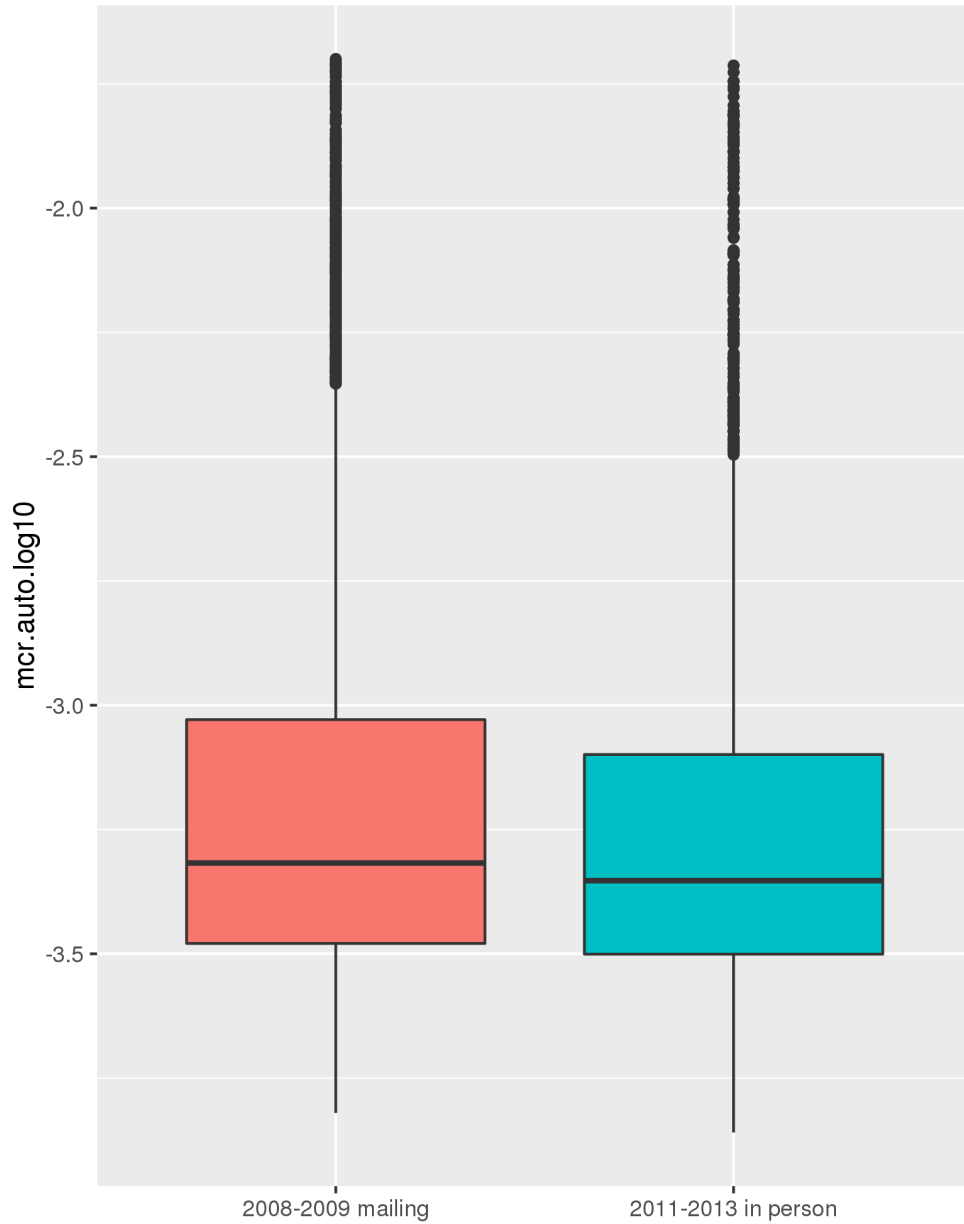


Figure 17: Boxplot of autosomal MCR for study samples categorized by genotyping plate. Red boxes indicate plates containing samples that failed in the first round of genotyping and were re-genotyped together (“redo” plates). The width of each box is proportional to the square root of sample size. Plates are ordered along the x-axis by plate name. All plates have low MCR.

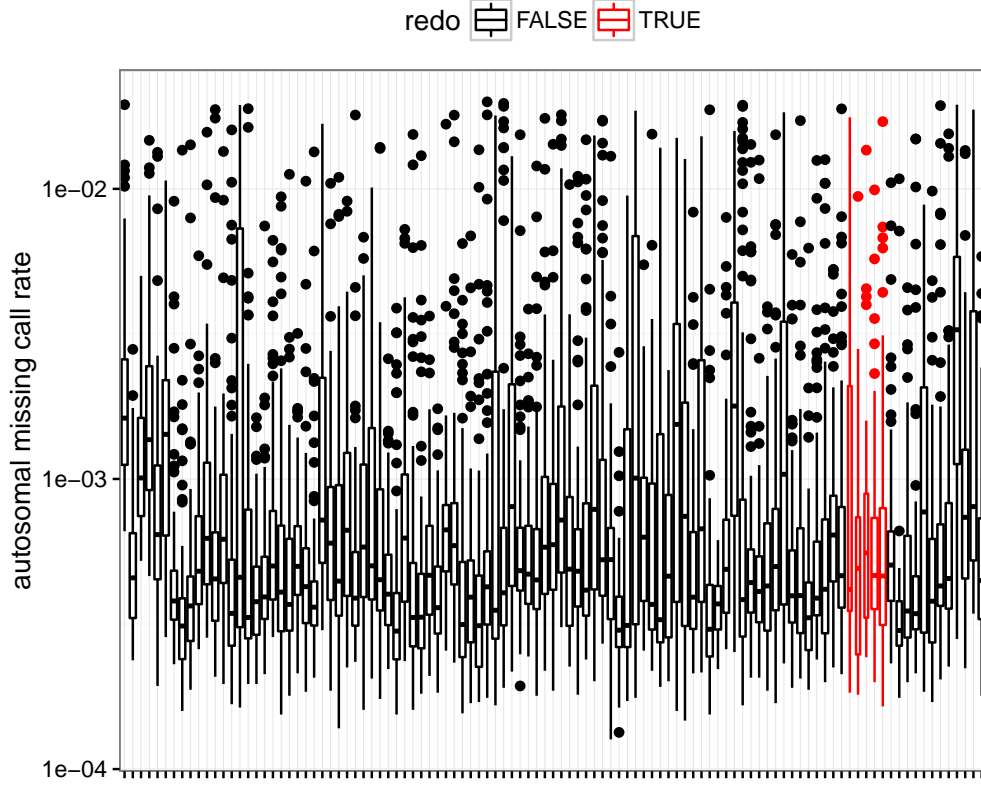
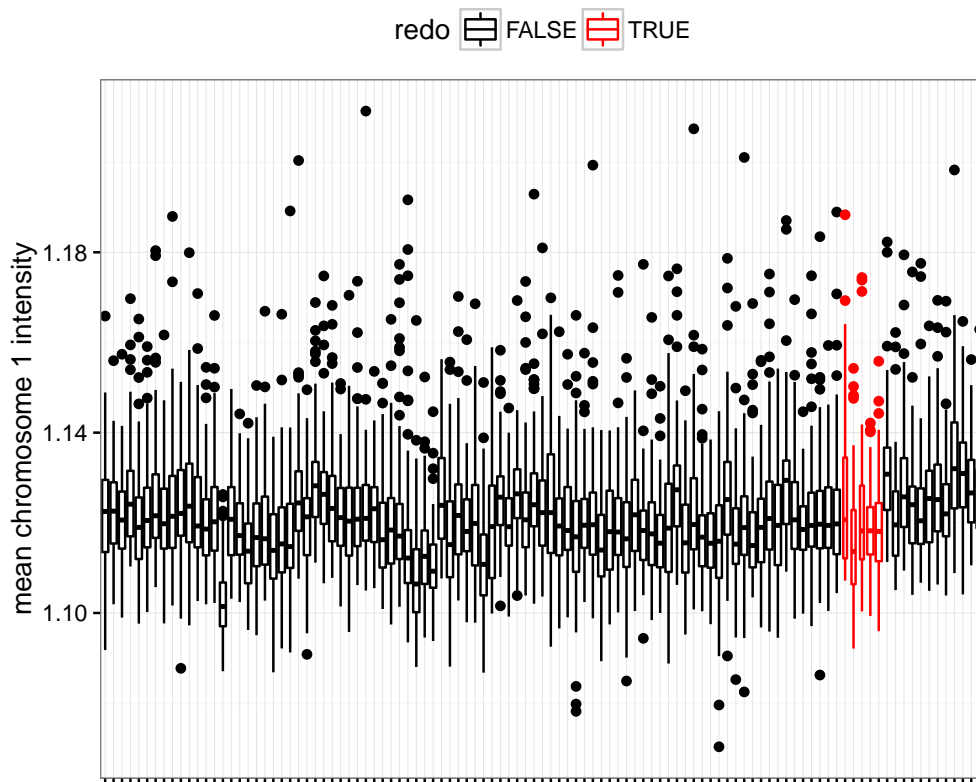
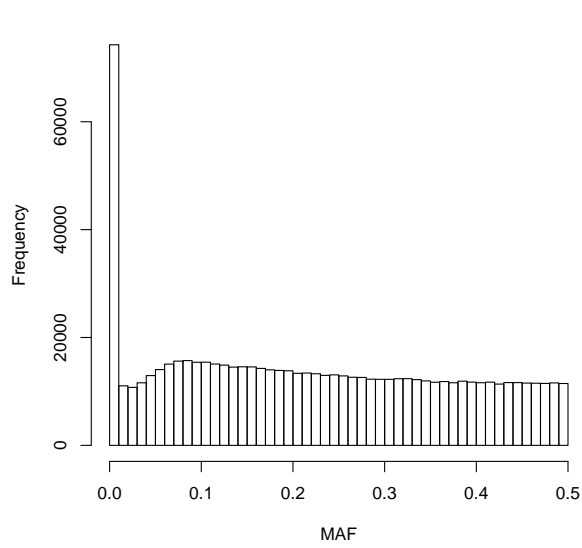
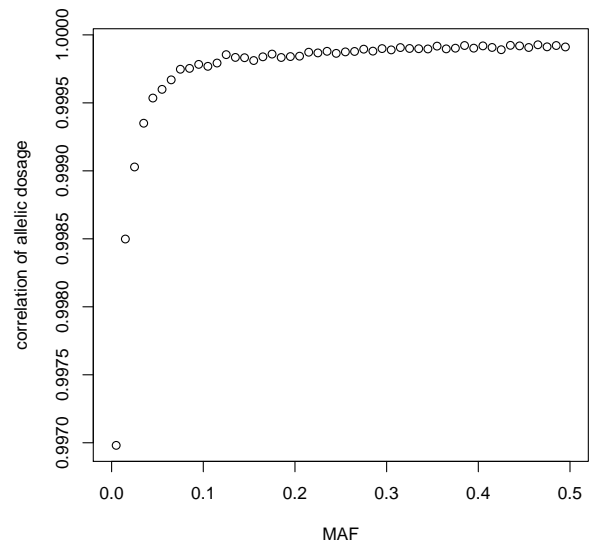


Figure 18: Boxplot of intensity for chromosome 1 probes. Samples are grouped by genotyping plate. Red boxes indicate plates containing samples that failed in the first round of genotyping and were re-genotyped together (“redo” plates). All plates have similar intensity profiles, suggesting there are no problematic plate/batch effects.

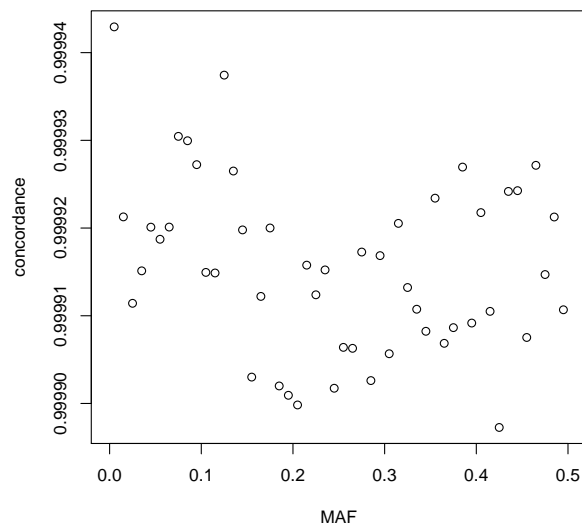




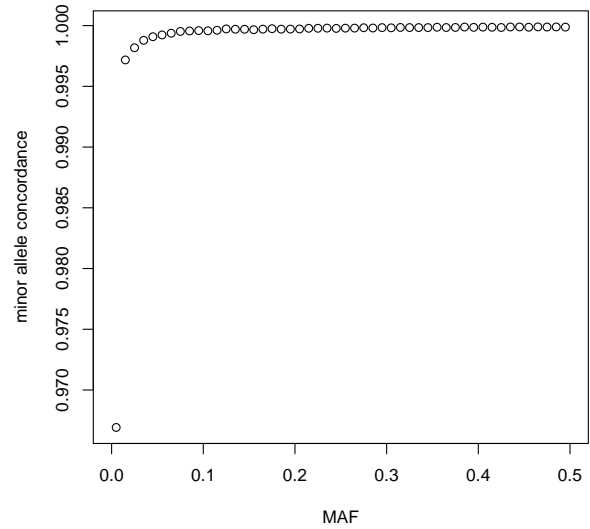
(a) Distribution of minor allele frequency.



(b) Correlation of allelic dosage.



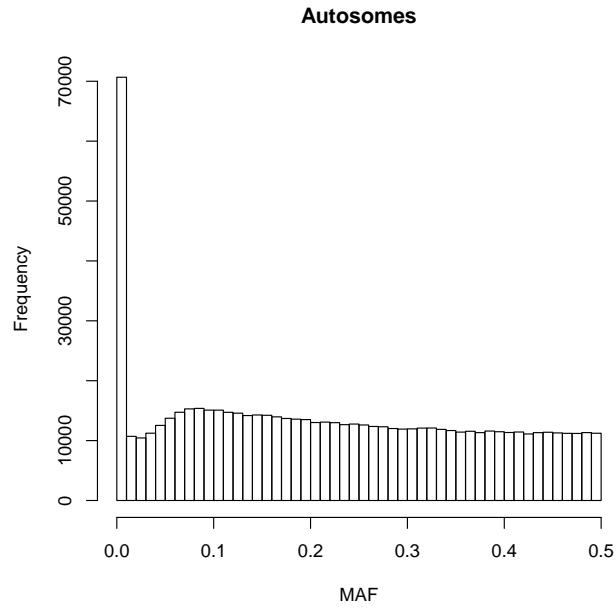
(c) Overall concordance.



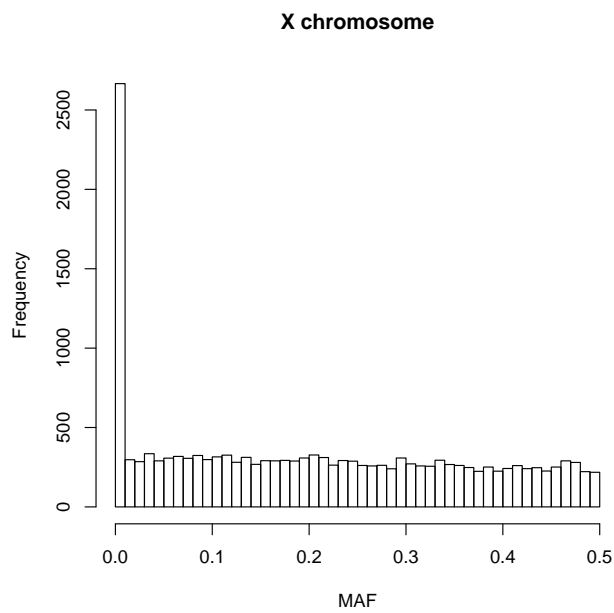
(d) Minor allele concordance.

Figure 19: Summary of concordance by SNP over 204 duplicate sample pairs, binned by minor allele frequency (MAF).

Figure 20: Minor allele frequency distribution across study participants.



(a) Autosomes



(b) X chromosome

Figure 21: The first two eigenvectors from the PCA performed in the set of homogeneous samples selected for HWE testing. The sample selection was made based on defining a two-dimensional ellipse of greatest density of self-reported “white” participants from the first two eigenvectors of the study-only PCA. Samples are color-coded according to self-identified race (AI/AN=American Indian/Alaska Native; multiple=More Than One Race; unknown=Unknown or Not Reported).

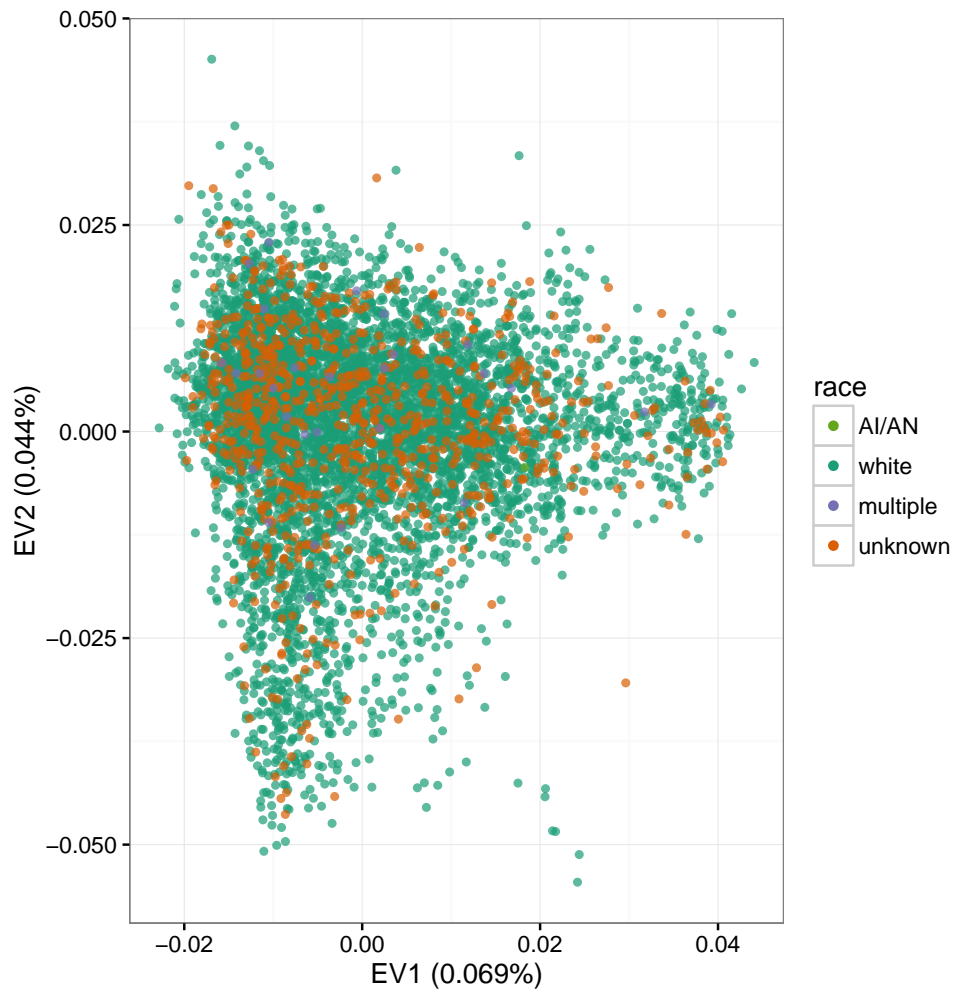


Figure 22: Distributions of estimated inbreeding coefficient for a random sample of autosomal SNPs, with black representing observed values calculated from the data and red representing values calculated from simulation assuming Hardy-Weinberg equilibrium. The potential values range from -1 to 1.

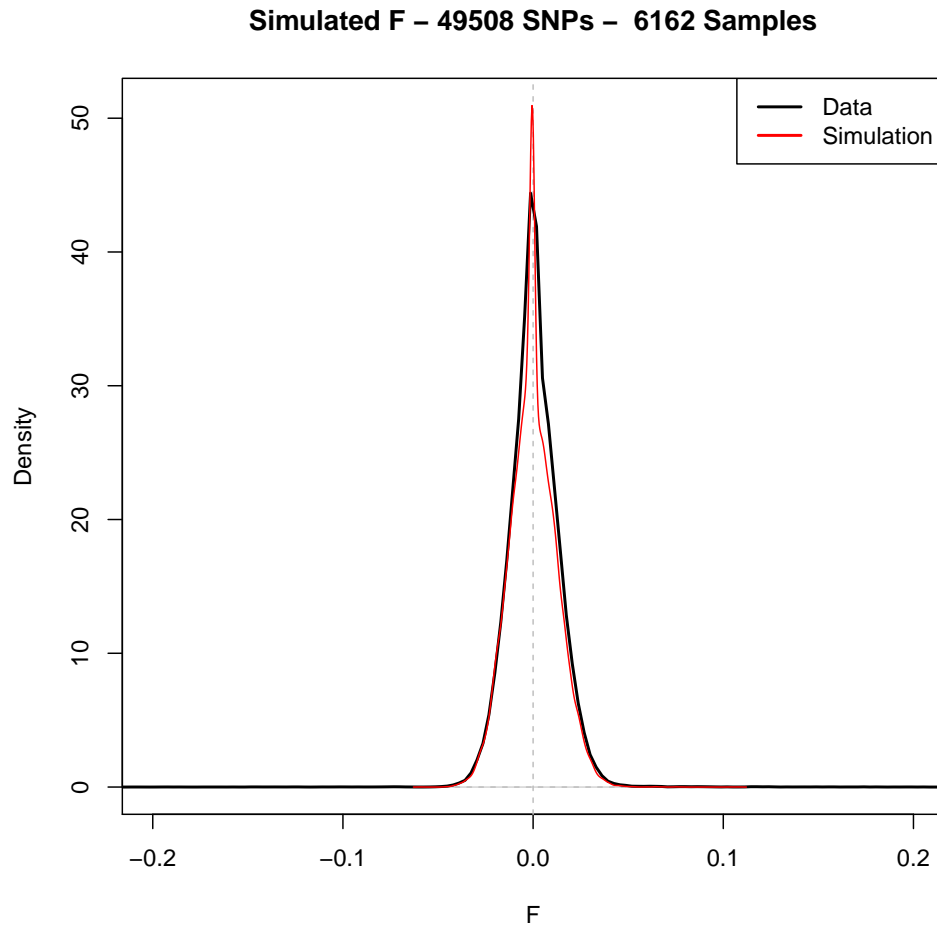


Figure 23: Quantile-quantile plots for preliminary association tests. QQ plots are provided after using no SNP filter, using composite filter, using composite filter plus MAF filter, and (bottom right plot) for SNPs satisfying the composite filter but MAF is lower than MAF filter threshold. The genomic inflation factor, lambda, is given below the x-axis label in each plot.

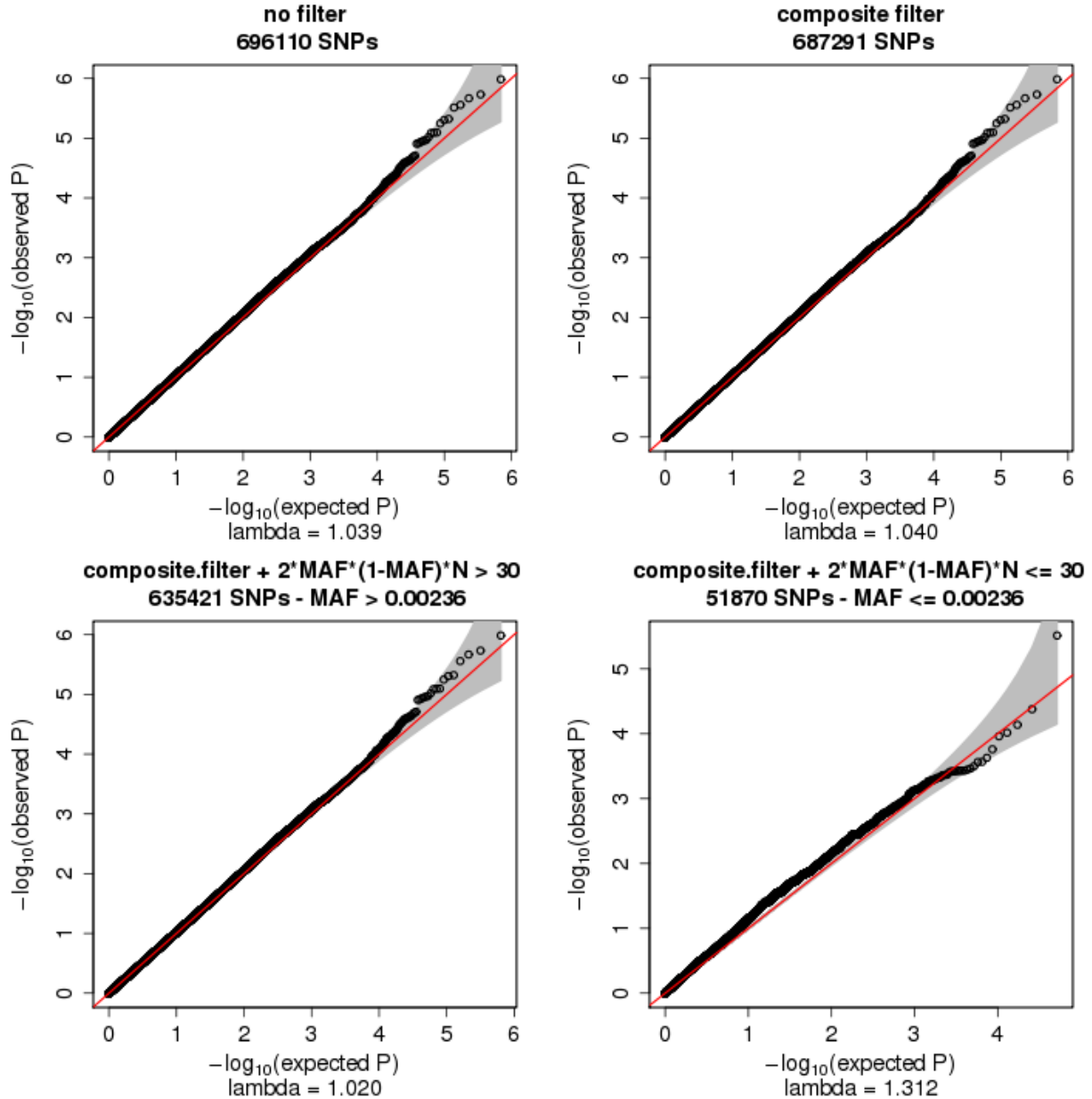


Figure 24: Manhattan plots for preliminary association tests.

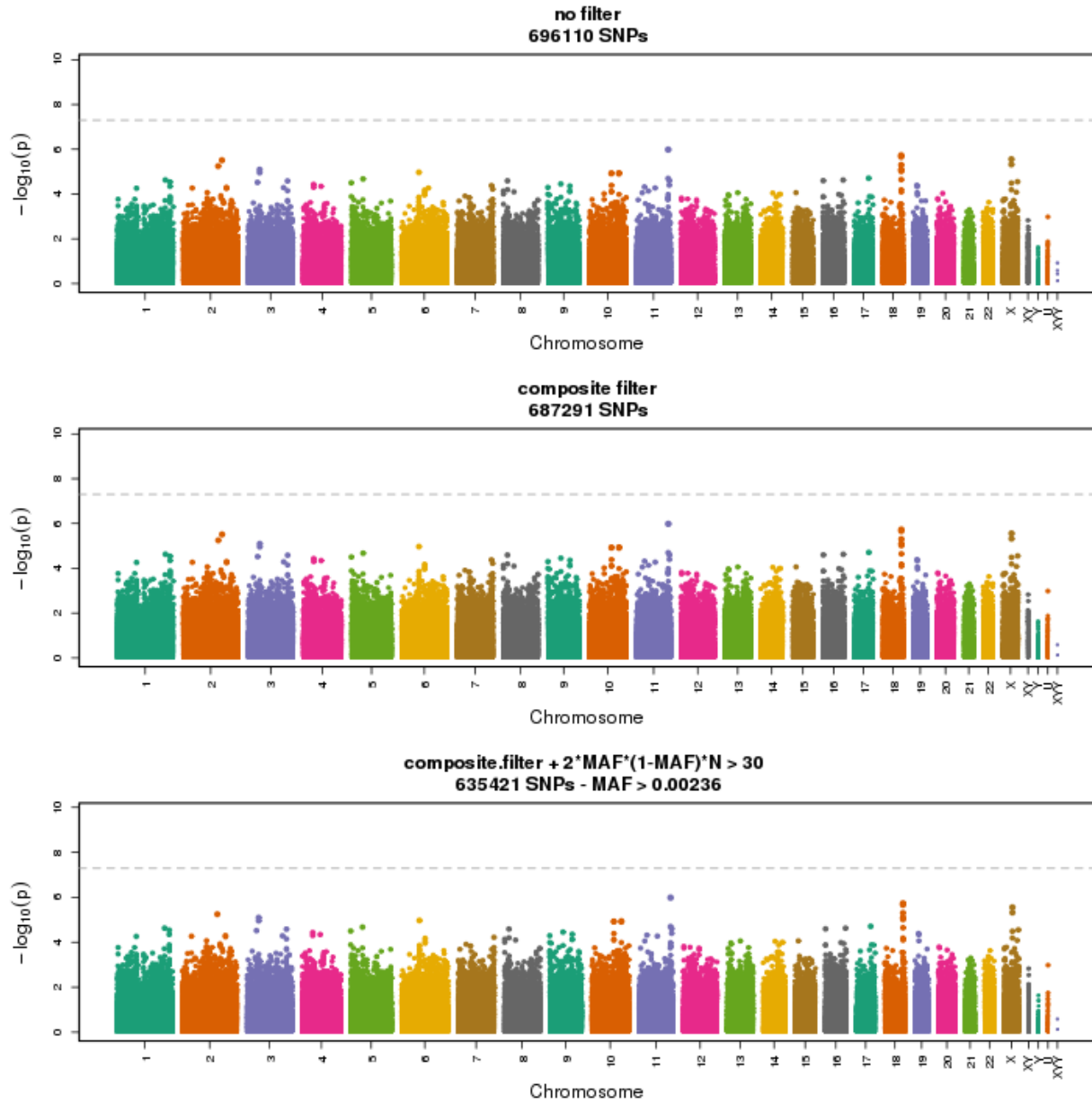


Figure 25: Genotype cluster plots for the top nine SNPs from the preliminary association test after applying the composite filter.

