Quality Control Report for Genotypic Data

University of Washington

September 22, 2014

Project: Efficiency of Bitter Blockers on Flavor Acceptance
Principal Investigator: Danielle R. Reed
Support: CIDR Contract # HHSN268201200008I
NIH Institute: NIDCD

Contents

List of Tables

List of Figures

1 Summary and recommendations for dbGaP users

A total of 140 study subjects were genotyped on the HumanOmni2.5m-8v1-1 array. These subjects were tested on up to four occasions for their perception of six bitter compounds and three bitter blockers as well as the anti-viral drug, Kaletra^{TM}. The median call rate is 99.9% and the error rate estimated from a single pair of study sample duplicate is 1.09×10^{-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 no 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

The goal of this project is to determine variation in human bitter perception and the effectiveness of bitter blockers in individuals of different genotypes. The project is based on the parent project titled "Efficacy of bitter taste blockers on flavor acceptance in pediatric population", which addresses how bitter taste perception affects medication compliance. As a part of the parent project, mothers and their children were recruited from the general population of metropolitan Philadelphia through print and internet media as well as from a database of past research participants. Adults (mothers) were evaluated in the laboratory for bitter taste perception for a variety of "Generally Recognized As Safe" (GRAS) bitter compounds using the general Labeled Magnitude Scale (gLMS) which allows subjects to rate sensory intensity [1]. Mothers were also evaluated for perception of the liquid formulation of the drug Kaletra^{TM}, a recommended 1st line treatment for infants with HIV infection, whose taste is rejected by some children and accepted well by others. DNA samples from mothers were genotyped as a part of this study to identify (1) novel genes associated with human bitter perception and (2) the ability of compounds to block bitterness.

3 Genotyping process

All the DNA study samples were derived from saliva using the DNA extraction method "oragene.prepIT.L2P". There were 5 HapMap control samples. Due to the small sample size, samples were genotyped in batches corresponding to 3 plates containing 46 samples each on average. Each batch contained an average of one (1) HapMap control and 1 or 2 duplicate study samples. Duplicate samples were not placed in the same batch. Samples were stratified by self-identified race.

The genotyping was performed at the Center for Inherited Disease Research (CIDR) using the Illumina HumanOmni2.5m-8v1-1 array (BPM annotation version B, genome build 37) and using the calling algorithm GenomeStudio version 2011.1, Genotyping Module version 1.9.4 and GenTrain version 1.0. The array consisted of a total of 2,391,739 SNPs. Note that earlier versions of Illumina annotations mis-annotated chromosome information for numerous SNPs designated as X or Y rather than as XY. These SNPs occur in pseudo-autosomal (PAR1, PAR2) regions or in the X-translocated region (XTR). The annotation was corrected prior to genotype calling.

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 University of Washington Genetics Coordinating Center (UWGCC), the study investigator's team and dbGaP. These data were analyzed by the analysis team at UWGCC 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* [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, 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 146 samples (including duplicates) were put into genotyping production all of which were successfully genotyped and passed CIDR’s QC process (Table 2). The subsequent QA process identified one (1) study sample with unusual karyotype. This sample was removed from the dataset. The set of scans to be posted include 140 study participants and 5 HapMap controls. The 140 study scans derive from 139 subjects and include 1 pair of duplicate scans (Table 3). The 5 HapMap control scans derive from 3 subjects, of which 2 are replicated two times. The study subjects occur as 126 singletons and 6 families of two or three members each. The study families were discovered during the analysis of relatedness (Section 8). The HapMap controls include 1 trio of ASW.

6 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 1 shows two distinct clusters with no apparent gender mis-annotation. (Note that the male cluster in this study is composed of HapMap control samples.)

Higher or lower than usual intensities or heterozygosities can be used to identify possible sex discrepancies or sex chromosome anomalies. Such samples are examined further by viewing BAF/LRR plots (see Section 7) Figure 1 identifies an individual with low X and autosomal heterozygosity. This individual was removed from the dataset due to an unusual karyotype after examining BAF/LRR plots.

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 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.

Only a single segmental anomaly was detected, which appears to be a segmental uniparental disomy with a very low level of mosaicism (i.e. primarily normal biparental disomy) and the genotype calls are not affected by this abnormality. Therefore, we did not filter the genotype calls for this individual.

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.

8 Relatedness

The relatedness between each pair of participants was evaluated by estimation of the kinship coeffcient(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.

IBD coefficients were estimated using 172,541 autosomal SNPs and the KING-robust procedure [9], but implemented in R using the package *SNPRelate* [3]. The SNPs 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) $> 5\%$ with all pairs of SNPs having $r^2 < 0.1$ in a sliding 10 Mb window. KING-robust was used because it is robust to population structure, which is needed for this mixture of multiple ethnic and ancestral groups (including African, European, Asian and Hispanic individuals). KING-robust provides estimates of the kinship coefficient and IBS0 (the fraction of SNPs that share no alleles), from which relationships can be inferred.

All study subjects were expected to be unrelated. All expected duplicates were identified and there were no unexpected duplicates. Figure 2 shows that there are several unexpected relationships. Because of the ambiguity in Degree 2 relationships, we were unable to specify a pedigree structure for the study subjects. Nevertheless, we defined families so that each family includes all pairs of subjects connected by a $KC > 0.08839 = 2^{-3.5}$, the lower boundary of inference for Degree 2 relationships (see Table 1 in [9]). This procedure resulted in 5 families of two members each, and one (1) family of three members. The IBD coefficient estimates for these families are provided in the file "Kinship_coefficient_table.csv." The PLINK files provided will have mother and father entries as NA for the subjects in these families. *For an analysis that assumes all participants are unrelated, we recommend selecting a maximal set of unrelated subjects from each family unit, using "unrelated" in "Sample_analysis.csv."*

9 Population structure

To investigate population structure, we use principal components analysis (PCA), essentially as described by Patterson et al. [10], but implemented in R (*SNPRelate* package). We use PCA for two purposes: to identify population group outliers and to provide sample eigenvectors as covariates in the statistical model used for association testing to adjust for possible population stratification.

We and others [11] 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 [11]. Therefore, the SNPs used for PCA were selected by LD pruning from an initial pool consisting of all autosomal SNPs with a missing call rate $< 5\%$ and minor allele frequency (MAF) $> 5\%$. In addition, the 2q21 (LCT), HLA, 8p23, and 17q21.31 regions were excluded from the initial pool. The LD pruning process selects SNPs from the initial pool with all pairs having $r^2 < 0.1$ in a sliding 10 Mb window.

Three PCA analyses were performed. The first PCA analysis combined (non-duplicated) study samples with an external set of HapMap controls to establish the ancestry orientation and to identify possible population group outliers. The second PCA analysis was performed on the set of all unrelated study subjects. The third PCA analysis was performed on a subset of unrelated subjects to determine a homogeneous ethnic group for use in Hardy Weinberg analysis.

(A) We performed PCA on all study subjects along with HapMap III subjects (Figure 3), using 101,466 pruned SNPs. (In addition to the pruning process described above, the initial pool of SNPs excluded any SNPs with a discordance between HapMap controls genotyped along with the study samples and those in the external HapMap data set.) Figure 3 is color-coded by self-identified race for study subjects and by population group for HapMap controls. Of the study subjects, 63% are African, 19% European, 17% Mixed (typically African and European ancestry) and less than 1% each of Hispanic and Asian descent. As expected, most of self-identified African and European study samples cluster around the HapMap ASW and CEU samples respectively. The single Hispanic study sample is located between the CEU and ASW HapMap samples and the Asian study sample is located around the GIH HapMap cluster as is often the case.

(B) We performed PCA on a set of 132 unrelated study subjects, using 163,695 pruned SNPs. The set of unrelated study subjects included all singletons and a maximal set of unrelated individuals from each family based on kinship coefficients (see discussion of families in Section 8). The logical variable "unrelated" identifies the set of unrelated subjects including HapMap controls and the logical vector "pca.study" identifies the set of unrelated study subjects used for this PCA in "Sample_analysis.csv". Figure 4 shows the plot of eigenvector 1 vs. eigenvector 2. The eigenvectors from this PCA were used in the association analyses.

To determine whether the LD-pruning effectively prevented the occurrence of small clusters of SNPs that are highly correlated with a specific eigenvector, we examine plots of the correlation of each SNP with each eigenvector. These plots are similar to GWAS "Manhattan" plots except that the Y-axis has the SNP-eigenvector correlation rather than an association test p-value. Figure 5 shows these plots for the first 8 eigenvectors. 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, the association of each eigenvector with the square root transformed traits from this study and the parallel co-ordinates plot. The scree plot(Figure 6) shows that the fraction of variance accounted for falls off dramatically after the second component. Association tests were performed for five traits namely, K.bitterK1, NA_P.bitter, NA_PB.bitter, SUC_B.sweet and MSG_U.bitter. None of the traits were found significant for any of the eigenvector after correcting for multiple testing. Figure 7 is a parallel-coordinates plot by self identified race for each subject across the first 12 eigenvectors. The parallel-coordinates plot helps visualize the relationship of self identified race to the population structure identified by the eigenvectors. Vertical lines represent eigenvectors and each piece-wise line traces eigenvector values for a given subject. For example, it can be seen that eigenvectors 1 and 2 separate out subjects of African ancestry (red lines) and European ancestry (blue lines). Because eigenvector 1 and 2 capture major continental ancestry these were used as covariates in the preliminary association tests described in Section 19.

The subjects to the left of the magenta line in Figure 4 show the selection of a more homogeneous set for use in Hardy Weinberg analysis. This set was selected as all unrelated subjects with eigenvector 1 less than zero. This cutoff selected 89 subjects.

(C) We performed PCA on the above set of 89 subjects. This set of subjects is identified by the logical variable "pca.afr" in "Sample_analysis.csv". Plotting eigenvector 1 vs eigenvector 2 in Figure 8 shows that this set of sample is homogenously distributed. We selected this sample set for performing HWE testing in Section 14.

10 Missing call rates

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

In this study, the two missing rates by sample are very similar, with median values of 0.0006 (*missing.e1*) and 0.0005 (*missing.e2*). Figure 9 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 0.000 and 98.93% of SNPs have a missing call rate $<$ 2%. All Y chromosome SNPs were dropped by CIDR since this is an all-female study. As a result the Y chromosome has a technical failure rate of 100%. The two males in this study are a pair of HapMap duplicates.

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 difference associated with the phenotype of interest can lead to spurious associations, since missingness is often nonrandom [12]. We tested for such a difference using linear regression of trait on missing call rate: $\sqrt[2]{Trait} \backsim log_{10}(miss.e1.auto)$ (autosomal missing call rate). Table 6 shows that none of the five traits were significantly associated with missing call rate ($p > 0.05$).

11 Batch effects

The samples were processed together in batches corresponding to 3 plates with 46 samples each on average. There is no significant variation among batches in log_{10} of the autosomal missing call rate ($p = 0.085$) and all plates have a low mean missing call rate (Figure 10).

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 (OR) for each SNP and each plate and then averaged these statistics over SNPs, using only study samples (which, in this study, are primarily of African ancestry). 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 11 shows the mean odds ratio (OR) compared with the number of samples per plate. There are no outlier plates, as plates with fewer samples are more likely to have allelic frequency differences compared to those with larger numbers of samples. We concluded that there are no problematic plate effects.

12 Duplicate sample discordance

Genotyping error rates can be estimated from duplicate discordance rates. The genotype at any SNP may be called correctly, or miscalled as either of the other two genotypes. If α and β are the two error rates, the probability that duplicate genotyping instances of the same participant will give a discordant genotype is $2[(1-\alpha-\beta)(\alpha+\beta)+\alpha\beta]$. When α and β are very small, this is approximately $2(\alpha+\beta)$ or twice the total error rate. Potentially, each true genotype has different error rates (i.e. three α and three β parameters), but here we assume they are the same. In this case, since the median discordance rate over all sample pairs is 2.18×10^{-5}, a rough estimate of the mean error rate is 1.09×10^{-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 7 shows these probabilities for $x = 0$ and $n = 1$. Here we chose $n = 1$ to correspond to the number of pairs of duplicate study samples.

We recommend a filter threshold of > 0 *discordant calls which retains* $>$ 99.8% *of SNPs with an error rate* $< 10^{-3}$*, while removing* $>$ 2% *of SNPs with an error rate* $> 10^{-2}$. This threshold eliminates 52 SNPs.

It would be desirable to eliminate a greater fraction of SNPs with error rates of 1e-2 while maintaining a high fraction of those with error rates of 1e-3, but having only a single pair of study sample duplicates severely limits the choices.

Figure 12 summarizes the concordance by SNP, binned by MAF. Given that there is only one pair of study sample duplicates, concordance by SNP is either 1 (gentoypes match) or 0 (they do not match); these values are averaged over all SNPs within each MAF bin. Figure 12a shows the number of SNPs in each MAF bin. Figure 12b 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.

13 Mendelian errors

Mendelian errors were not computed for this study since the study had only 1 HapMap trio. Typically, 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.

14 Hardy-Weinberg equilibrium

We calculated an exact test of Hardy-Weinberg equilibrium (HWE) using study subjects who are (1) unrelated, (2) have missing call rate $< 2\%$, (3) "pca.afr" subjects described in Section 9. A total of 89 samples are selected out of which 81 are of African ethinicity and 8 have mixed ethnicity. The logical variable "pca.afr" in "Sample_analysis.csv" indicates this sample selection. Figure 13 shows quantile-quantile (QQ) plots for the HWE tests. The autosomes deviate strongly from expectation at a p-value of about 0.001, while the X chromosome SNPs are reasonably consistent with expectation throughout the range of p-values. 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 14 shows a comparison of the observed distribution of the inbreeding coefficient estimates (for a random sample of 48,654 autosomal SNPs) with a simulated distribution of inbreeding coefficient estimates for the same set of SNPs under the assumption of Hardy-Weinberg equilibrium. The distributions are very similar. 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 0.0001 and 0.01 , *we suggest using a filter threshold of* $p = 0.0001$ *because examination of cluster plots reveals good plots for many assays with p-values* > 0.0001. This threshold is rather subjective, but we are reluctant to recommend a higher threshold that would eliminate many good SNP assays.

15 Minor allele frequency

Figure 15 shows the distribution of minor allele frequency (MAF) for all study subjects. The percentage of all SNPs with MAF $< 1\%$ is 15% for the autosomes and 13% for the X chromosome.

16 Duplicate SNP probes

The HumanOmni2.5m-8v1-1 array has 9,534 sets of SNPs that occur as apparent replicates, as indicated by identical genomic map positions within each set ("positional duplicates"). These all occur as pairs. Concordance of genotype calls across study samples for each pair of SNPs with the same map position was calculated. A high level of concordance indicates that these SNPs assay the same variant. To determine a suitable cut-off for concordance, we calculated the probability of having $> x$ discordant calls over 140 study samples, given assumed error rates. We chose > 0 discordances, for which the probability is 0.24 with error rate of 0.001 and 0.94 with error rate 0.01. Pairs with < 1 discordances are considered to assay the same SNP and one member of each pair is labeled as "redundant" in "SNP_analysis.csv" (the one with higher missing call rate). Pairs with > 0 discordances may be assaying different SNPs and are flagged as discordant by "dup.pos.disc" = TRUE in "SNP_analysis.csv". There were 9,367 redundant SNPs and 167 positions flagging 334 discordant duplicates.

17 Sample exclusion and filtering summary

As discussed in Section 5, genotyping was attempted for a total of 146 samples and all of them passed CIDR's QC process (Section 2). The subsequent QA process identified no quality or identity issues, but one sample to be excluded because of an unusual karyotype. Therefore, 140 study sample scans will be posted on dbGaP, along with 5 HapMap controls (a total 145 scans will 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%. However, no such anomalies or samples were found in this study. Therefore, no quality-based sample filters are indicated. 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).

18 SNP filter summary

Table 1 summarizes SNP failures applied by CIDR prior to data release and a set of additional filters suggested for removing assays of low quality or informativeness. The suggested quality filter (from rows 2 - 5) and composite filter (from rows 2 - 7) are provided as logical variables in the "SNP_analysis.csv" file, which also has the individual quality metrics so that the user can apply alternative thresholds. The quality filters (rows 2 - 5) remove 1.39% of the 2,391,739 SNP assays attempted and the composite filter (rows 2 - 8, also excluding uninformative redundant SNPs and monomorphic SNPs) removes 10.68% of the SNP assays.

In addition to the composite filter, we also 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 16.68% of the SNP assays attempted. *Regardless of what filters are applied to association test results, it is highly recommended to view SNP cluster plots for any SNP of interest.*

19 Preliminary association tests

Preliminary association tests were performed using linear regression for five of the 95 quantitative traits. Table 8 gives the description of the five traits provided as variables "K.bitterK1", "NA_P.bitter","NA_PB.bitter", "SUC_B.sweet", "MSG_U.bitter" in "Sample_analysis.csv" file. Square root transformed values for each trait were used in all association tests. 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. Autosomal SNPs were coded as 0, 1 or 2 for BB, AB and AA).

The subjects used in association tests consisted of unrelated subjects assayed for the trait. The variables "K.bitterK1.assoc", "NA_P.bitter.assoc","NA_PB.bitter.assoc","SUC_B.sweet.assoc", "MSG_U.bitter.assoc" in "Sample_analysis.csv" are logical indicators for the selection of subjects for the trait variables "K.bitterK1", "NA_P.bitter","NA_PB.bitter","SUC_B.sweet", "MSG_U.bitter" repectively. See Table 8 for the number of subjects used for each association test. The sample sizes for each trait are considerably less than the total of 140 subjects because the measurements where done in sub-studies consisting of different subsets of subjects. See Section 8 for a discussion of the determination of unrelated subjects. All of the samples used for the association tests had missing call rate $< 2\%$.

As discussed in Section 9, eigenvectors 1 and 2 were selected for use as covariates for all five traits. Note this is an all-female study and hence sex was not included as a covariate. The final association model used for each trait is: $\sqrt[2]{Trait} \backsim$ EV1 + EV2 + SNP. For conciseness we have included association test results for only two traits namely "NA_P.bitter" and "SUC_B.sweet".

For the trait variable "NA_P.bitter", Figure 16 shows the QQ plots for likelihood ratio tests of the SNP effect. 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 an MAF filter of MAF > 0.151 for "NA_P.bitter". Figure 16 shows that there is low genomic inflation (lambda = 1.020 to 1.026) for the majority of SNPs, but some inflation for low MAF SNPs (lambda = 1.033) for "NA_P.bitter". In addition, we found no evidence of isolated SNPs with low p-values or other evidence of false positives. The corresponding Manhattan plots for "NA_P.bitter" are shown in Figure 17. The plots show two SNPs with p-values less than the genome-wide threshold of 5×10^{-8} on chromosome 7. These SNPs with rsIDS rs10246939 and rs713598 (Illumina SNP "Name" kgp5899594) have been reported to be significantly associated with bitter taste perception previously and lie within the bitter receptor gene TAS2R38 [13]. Cluster plots for the two significant SNPs and the next top 7 hits for "NA_P.bitter" are shown in Figure 18. Most of the SNPs in these plots show good clustering.

For the "SUC_B.sweet" the effective sample size filter yeilds a MAF of > 0.189. The QQ plots for "SUC_B.sweet" in Figure 19 including low MAF SNPs show an abnormal pattern that appears to be due to an unusual distribution of the trait in combination with low MAF. The QQ plot filtered for higher MAF conforms well with an expected null distribution. There is low genomic inflation (lambda = 1.004 to 1.025) for the majority of SNPs and some deflation for low MAF snps SNPs (lambda = 0.9897). The corresponding Manhattan plots for "SUC_B.sweet" are shown in Figure 20. Cluster plots for the top 9 hits for "SUC_B.sweet" are shown in Figure 21. Most of the SNPs in these plots show good clustering.

We suggest that users check trait distributions and try different transformations for traits with abnormal QQ plots.

Appendix

A Project participants

Monell Chemical Senses Caenter
Danielle R. Reed, Julie Mennella and Cailu Lin

Center for Inherited Disease Research, Johns Hopkins University
Kim Doheny, Jane Romm, Hua Ling and Elizabeth Pugh

Genetics Coordinating Center, Department of Biostatistics, University of Washington
Deepti P Jain, Cecelia Laurie, Cathy Laurie and Bruce Weir

dbGaP-NCBI, National Institutes of Health
Nataliya Sharopova

References

[1] Mennella JA, Reed DR, Roberts KM, Mathew PS, and Mansfield CJ. Age-related differences in bitter taste and efficacy of bitter blockers. *PLoS ONE*, 9(7):e103107, 2014.

[2] 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.

[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] 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.

[13] Reed DR, Zhu G, Breslin PA, Duke FF, Henders AK, et al. The perception of quinine taste intensity is associated with common genetic variants in a bitter receptor cluster on chromosome 12. *Hum Mol Genet*, 19:4278–4285, 2010.

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 - 7 comprise the composite.filter with rows 2 - 5 being quality metrics and rows 7 and 8 being informativeness metrics.

Row.No	Filter	SNPs.lost	SNPs.kept
1	None(all SNP probes)	NA	2,391,739
2	CIDR technical filters	7,446	2,384,293
3	Missing call rate >=2%	25,600	2,358,693
4	>0 discordant calls in 1 study duplicate	39	2,358,654
5	HWE p-value <10^-4	236	2,358,418
6	MAF = 0	213,192	2,145,226
7	positional duplicates	8,925	2,136,301
8	MAF <0.01	143,521	1,992,780
	Percent of SNPs lost due to quality filters (rows 2-5)	1.39	
	Percent of SNPs lost due to composite filter (rows 2-7)	10.68	
	Percent of SNPs lost due to composite filter and MAF (rows 2-8)	16.68	

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

Category	Study	HapMap	Both
DNA samples into genotyping production	141	5	146
Failed samples	0	0	0
Scans released by genotyping center	141	5	146
Post-release QC failure	0	0	0
Scans with unresolved identity issues	0	0	0
Scans with other issues	-1	0	-1
Scans to post on dbGaP	140	5	145

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

Category	Study	HapMap	Both
Scans to post on dbGaP	140	5	145
Subjects	139	3	142
Replicated subjects	1	2	3
Families (N>1)	6	1	7
Singletons	126	0	126

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. Note that all Y chromosome SNPs were dropped by CIDR since this is an all-female study. As a result the Y chromosome has a technical failure rate of 100%. The two males in this study are duplicates of a HapMap subject.

	A	M	U	X	XY	Y
number of probes	2,326,391	239	7,576	51,908	3,587	2,038
SNP tech failures	0.00221	0.02929	0.02508	0.00108	0.00279	1.00000
missing>0.05	0.00221	0.02929	0.02508	0.00108	0.00279	1.00000

Table 6: Summary of linear regression of trait on missing call rate: $\sqrt[2]{Trait} \backsim log_{10}(miss.e1.auto)$ (autosomal missing call rate).

	Phenotype	p-value
1	K.bitterK1	0.23
2	NA_P.bitter	0.40
3	NA_PB.bitter	0.33
4	SUC_B.sweet	0.90
5	MSG_U.bitter	0.19

Table 7: Probability of observing more than the given number of discordant calls in a pair 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 > 0 discordant calls.

	Assumed error rate				
# discordant calls	1.0e-05	1.0e-04	1.0e-3	1.0e-2	# SNPs
>0	2e-05	0.0002	0.002	0.02	52

Table 8: Description of the phenotypic trait variables and their sub-study classification. The column "Sample number" indicates the number of samples used for preliminary association tests. PROP = Propylthiouracil, NaGlu = Sodium Gluconate, MSG = Monosodium Glutamate, gLMS = General Labeled Magnitude Scale

	Trait variable	Sub-study	Trait description	Sample number
1	K.bitterK1	Kaletra	gLMS bitter rating for Kaletra	83
2	NA_P.bitter	NaGlu blocker	gLMS bitter rating for PROP	117
3	NA_PB.bitter	NaGlu blocker	gLMS bitter rating for PROP + NaGlu	117
4	SUC_B.sweet	Sucrose blocker	gLMS sweet rating for sucrose	98
5	MSG_U.bitter	MSG blocker	gLMS bitter rating for urea	104

Figure 1: The X and Y chromosome 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. The two males are a pair of HapMap duplicates.

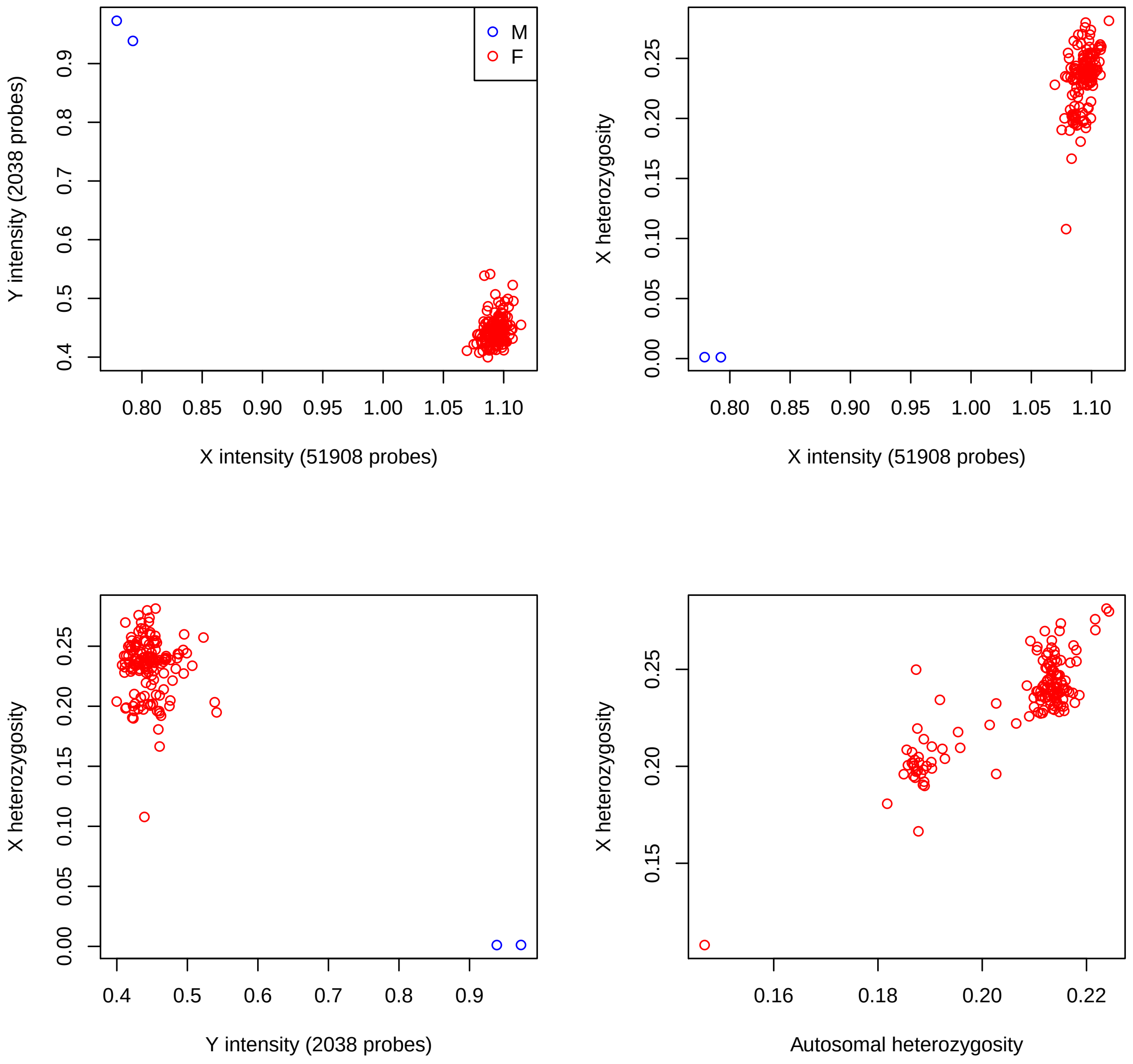


Figure 2: IBD coefficients to estimate relatedness. Each point represents a pair of samples. This plot shows 11 study pairs of participants with an estimated $KC > 1/32$, color-coded by inferred relationships. Gray dashed horizontal lines show boundaries for KC values for inferring varying degrees of relatedness. The first and second (from the top) 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 degree or higher (See Table 1 in [9]). The vertical dashed gray line represents a guideline for designating PO pairs or duplicates (whose IBS0 is theoretically 0). In the legend, "PO" = parent-offspring, "Dup" = duplicates, "FS" = full siblings, "Deg2"=Degree 2 relationships (such as half-siblings/avuncular/grandparent/grandchild), "Deg3"=Degree 3 relationships (such as first cousin), and "Unrel" = unrelated samples. All duplicates shown are expected duplicates. However, the other relationships are unexpected since all subjects were expected to be unrelated.

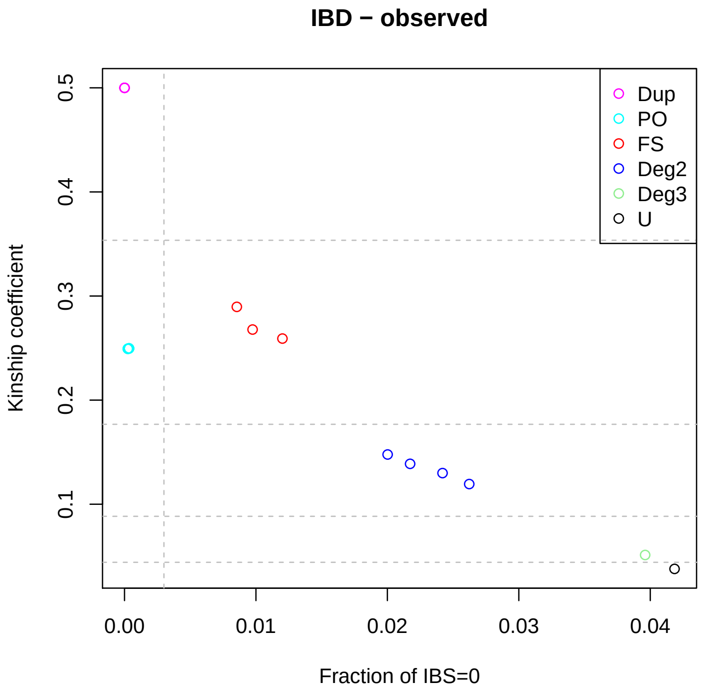
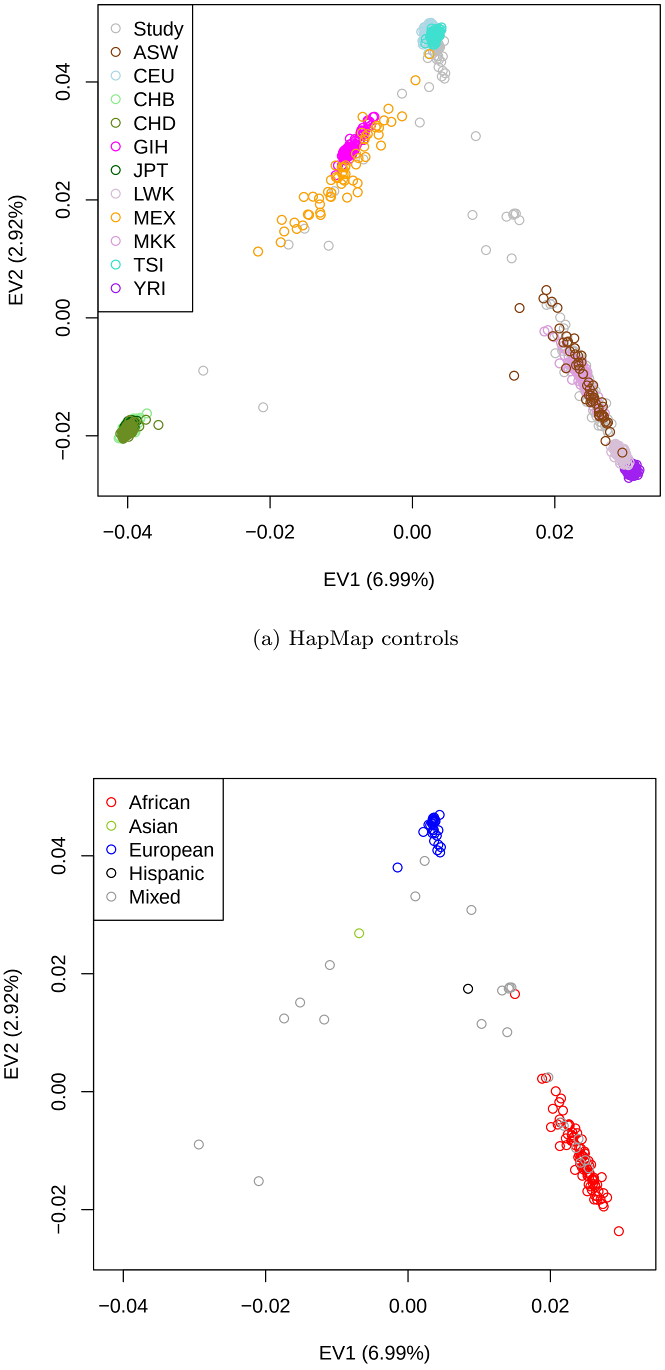


Figure 3: Principal component analysis of 139 study subjects with 1,202 HapMap controls. Separate plots (on the same scale) of HapMap controls and study subjects are provided for ease of comparison. Color-coding is according to self-identified race for study subjects and population group for HapMap controls. Axis labels indicate the percentage of variance explained by each eigenvector.



(a) HapMap controls

(b) Study subjects

Figure 4: Principal component analysis of 132 unrelated study subjects without HapMap controls. Color-coding is according to self-identified race. These subjects are identified by the logical variable ”pca.study” in ”Sample analysis.csv”. Axis labels indicate the percentage of variance explained by each eigenvector. Subjects to the left of the dotted magenta line (89 subjects) are selected for HWE analysis. These subjects are identified by the logical variable ”pca.afr” in ”Sample analysis.csv”.

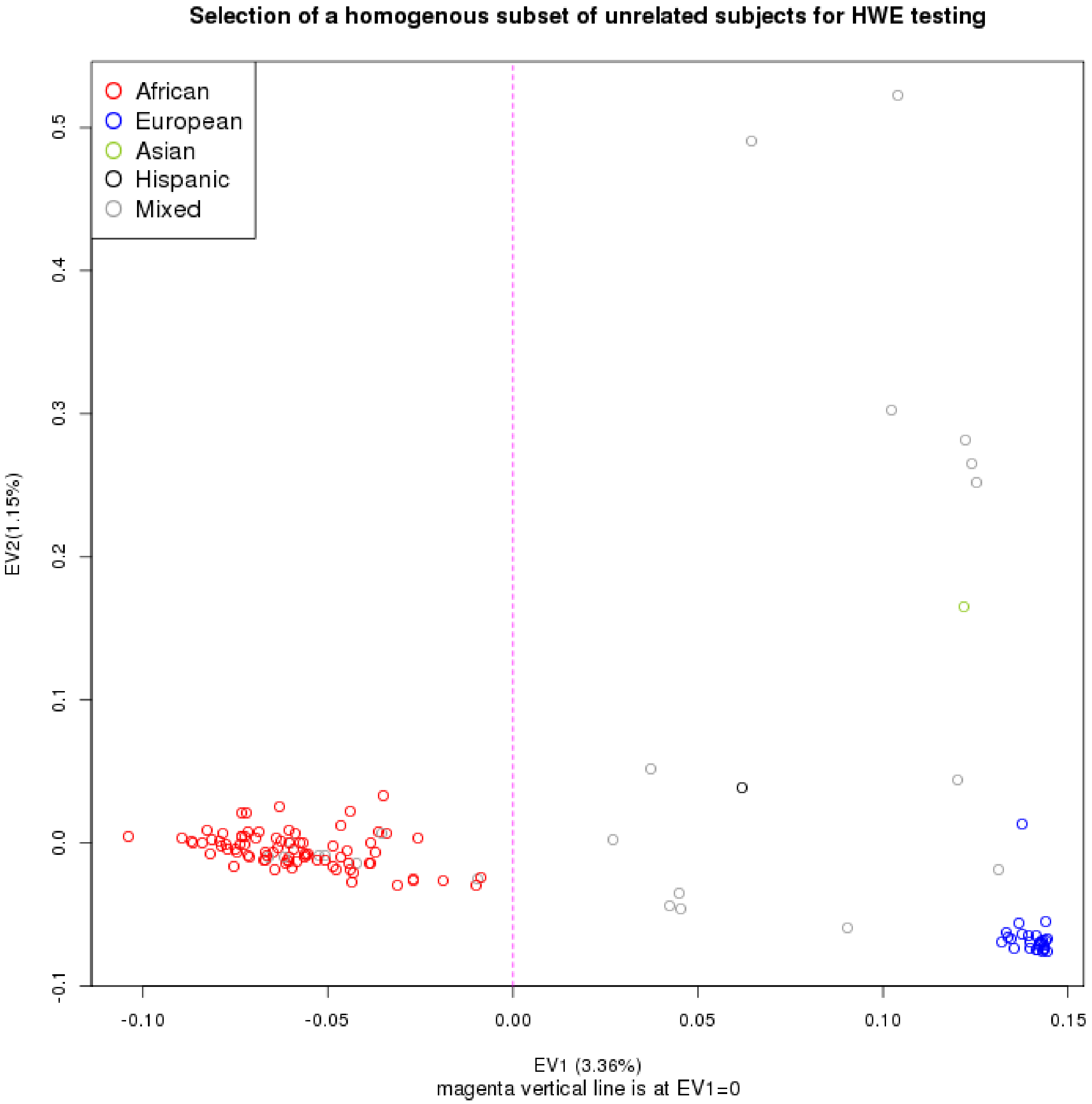
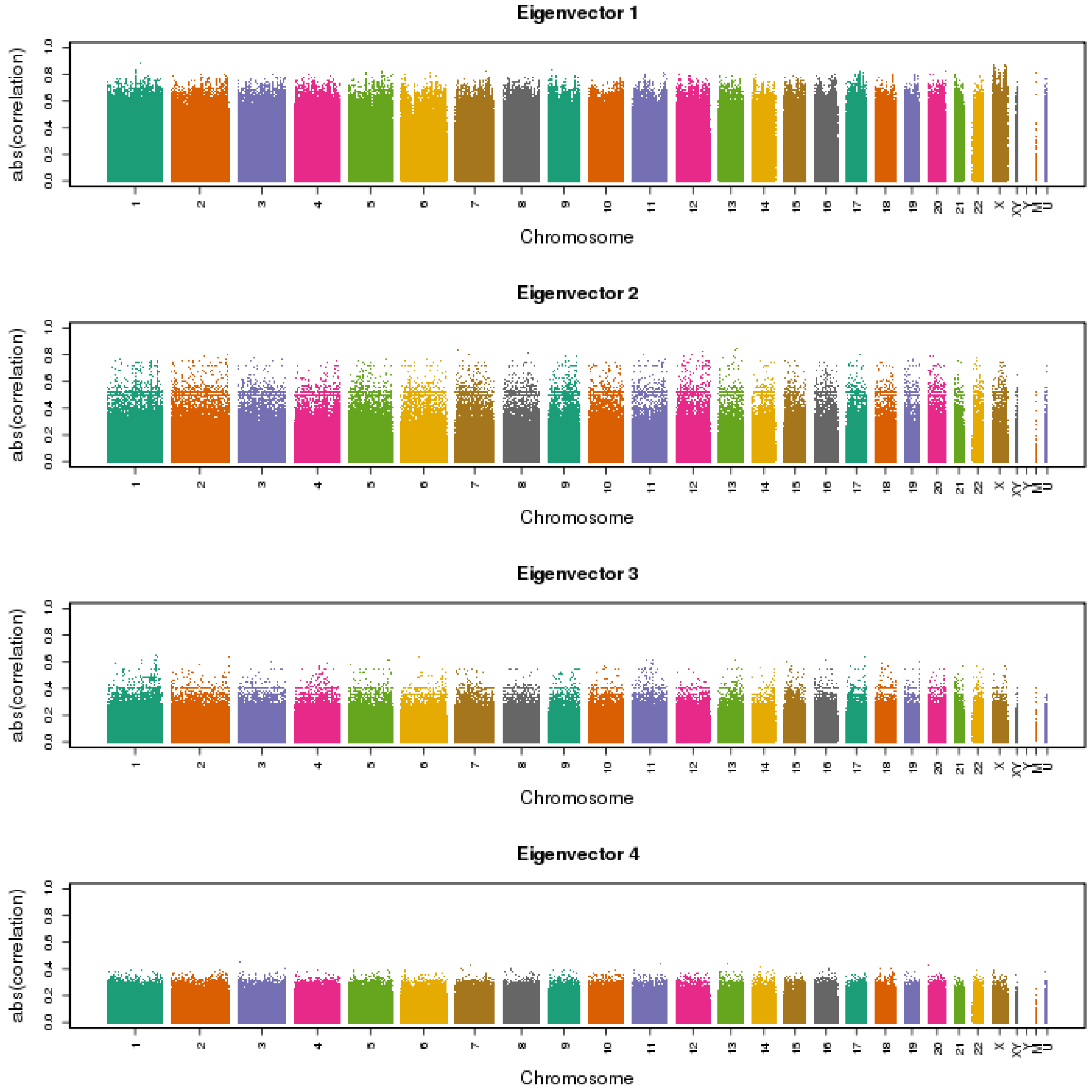


Figure 5: SNP position versus correlation between SNP genotype (0, 1 or 2) and each of the first 8 eigenvectors. These eigenvectors are from the PCA of ”pca.study” unrelated study subjects.



Eigenvector 1

Eigenvector 2

Eigenvector 3

Eigenvector 4

Figure 5: Continued.

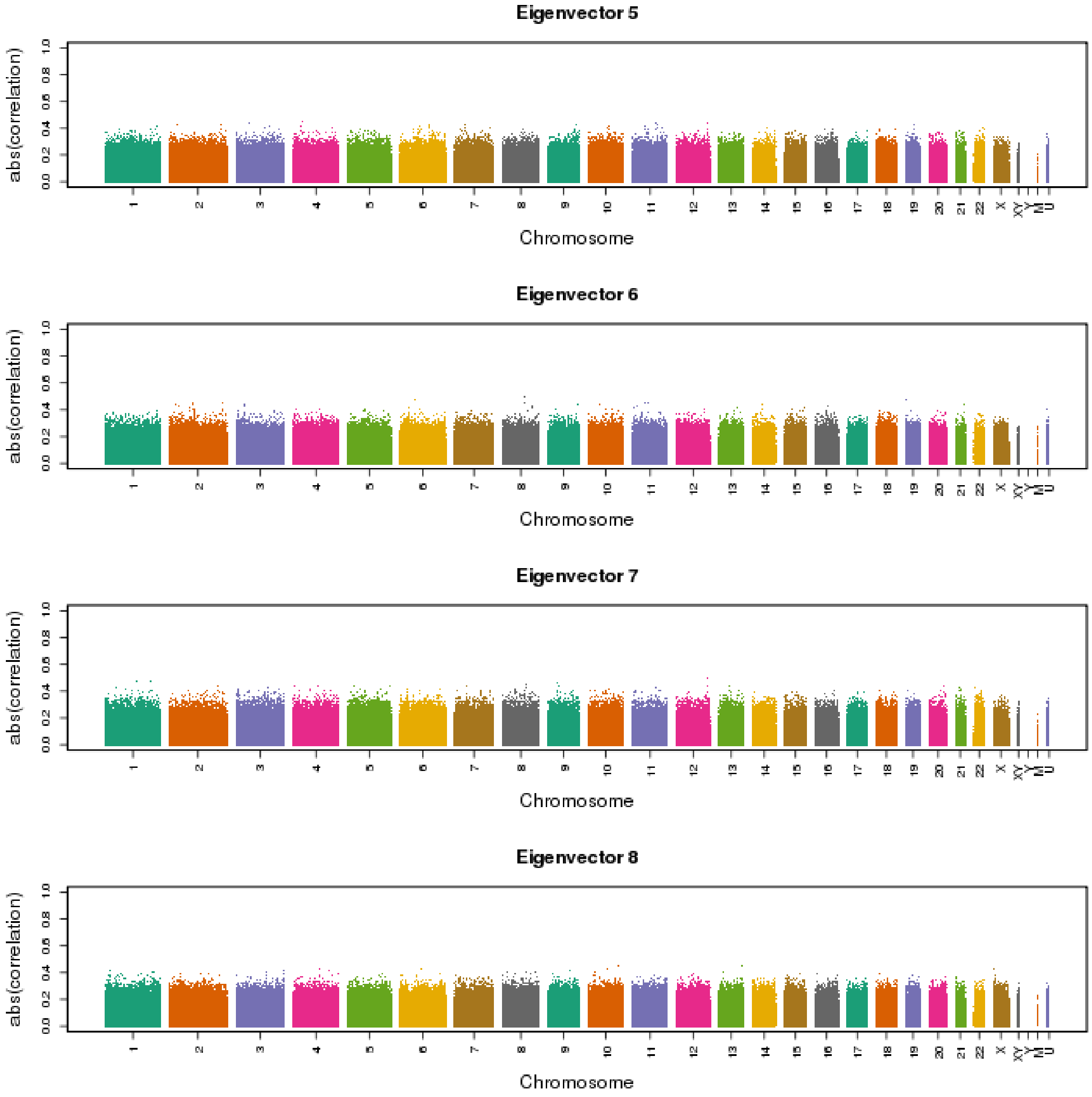


Figure 6: Scree plot for PCA shown in Figure 4.

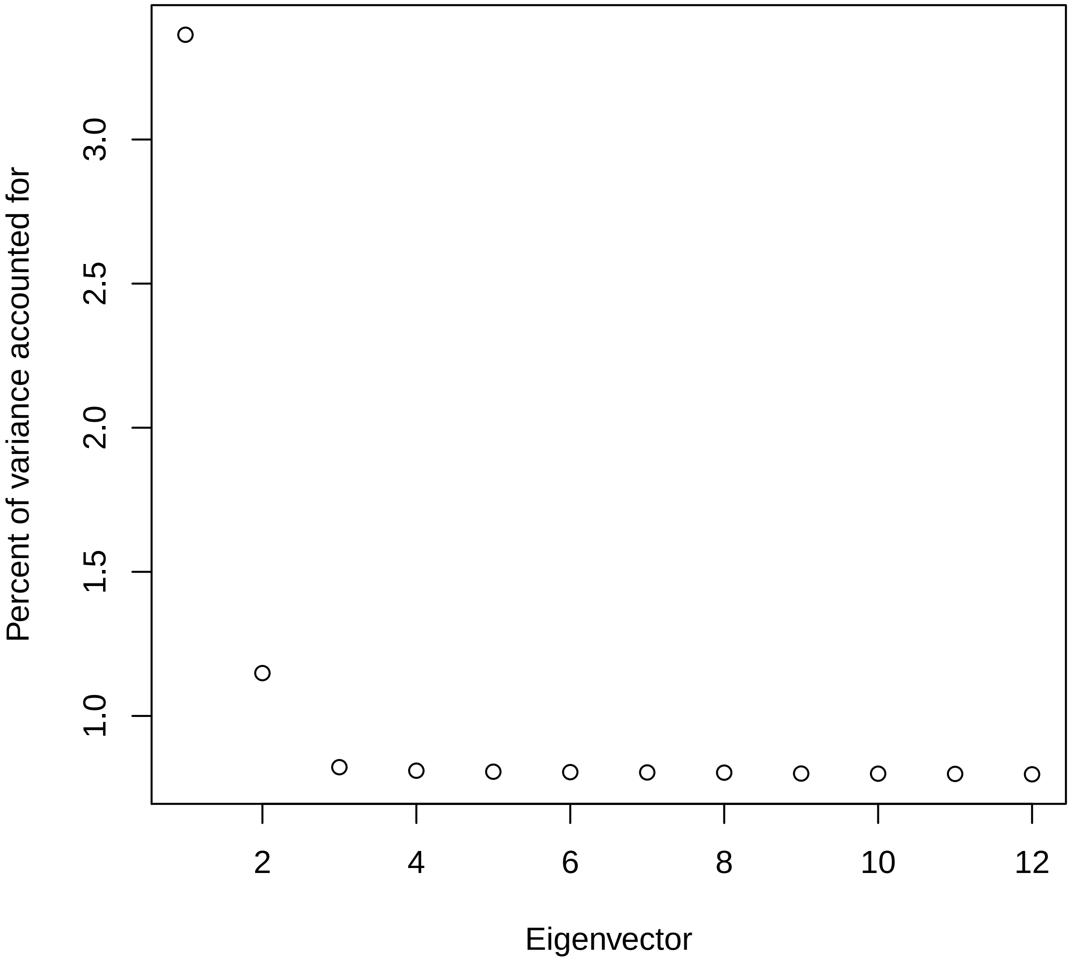


Figure 7: Parallel coordinates plot for visualization of relationship of PCA eigenvector structure with self identified race for the first 12 eigenvectors. Vertical lines represent eigenvectors and each piece-wise line traces eigenvector values for a given subject. These eigenvectors are from the PCA of ”pca.study” unrelated study subjects. Color-coding is according to self identified race.

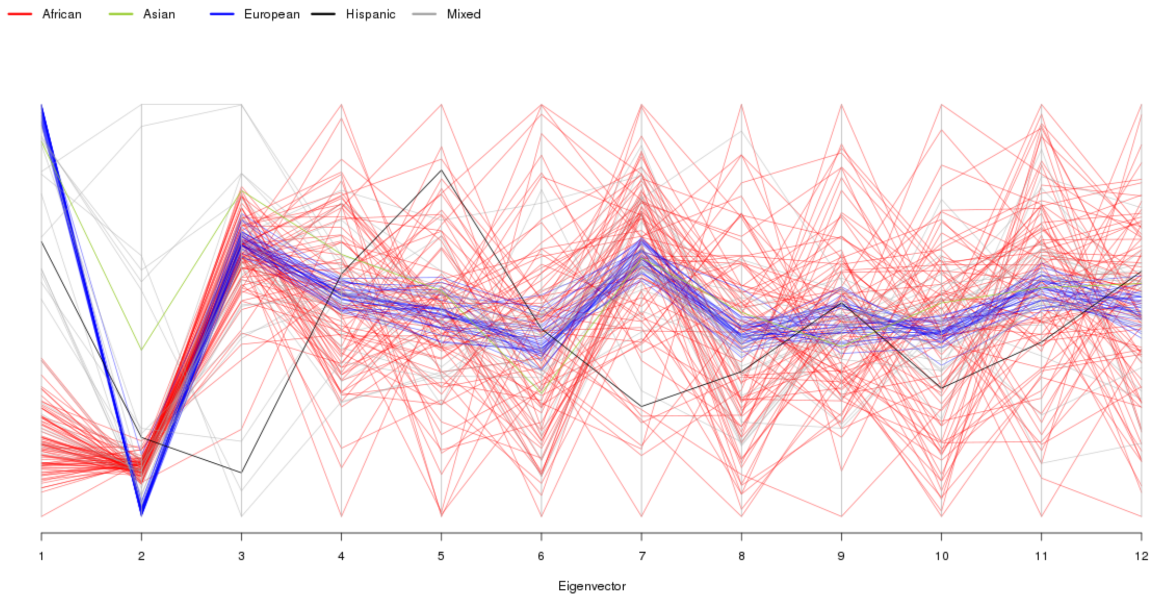


Figure 8: Principal component analysis of 89 unrelated study subjects (identified by the logical variable "pca.afr" in "Sample_analysis.csv"). Color-coding is according to self-identified race. Axis labels indicate the percentage of variance explained by each eigenvector.

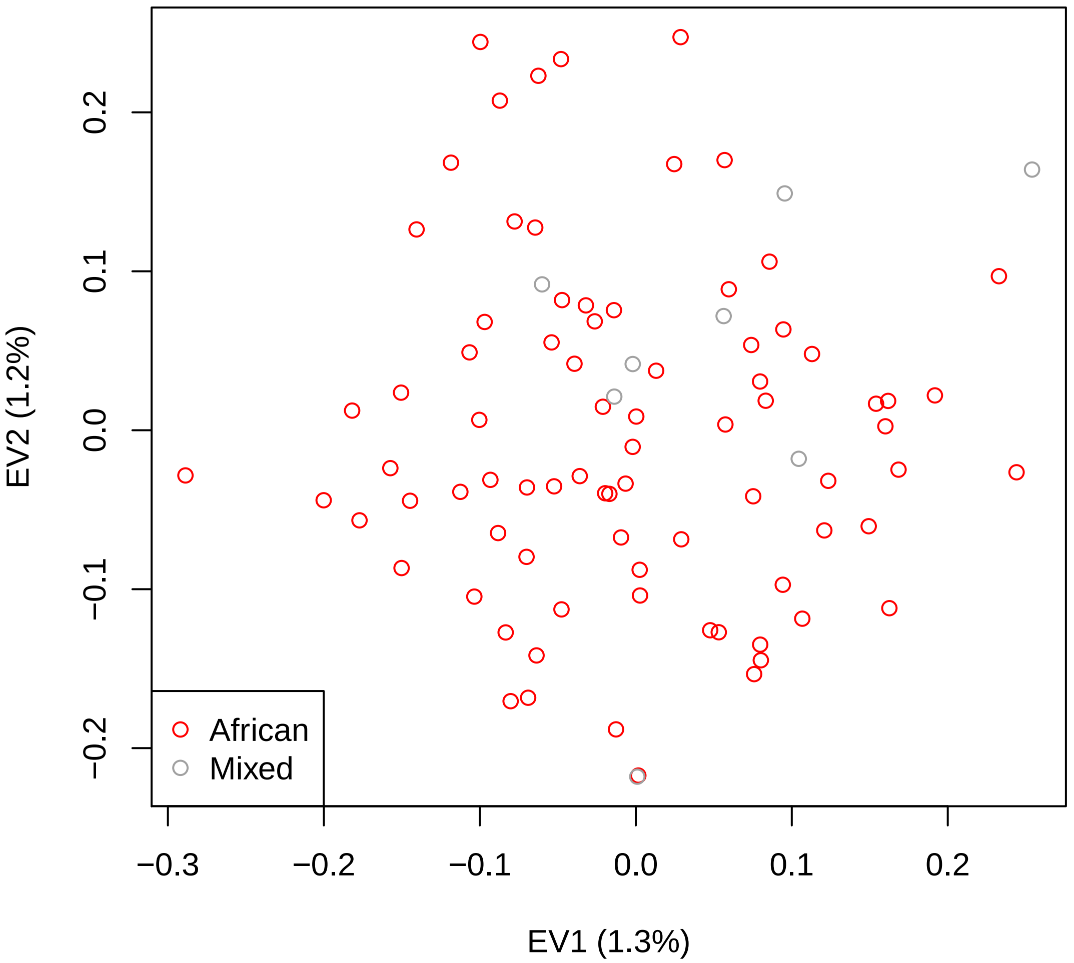


Figure 9: Histogram of the missing call rate per sample (*missing.e1*).

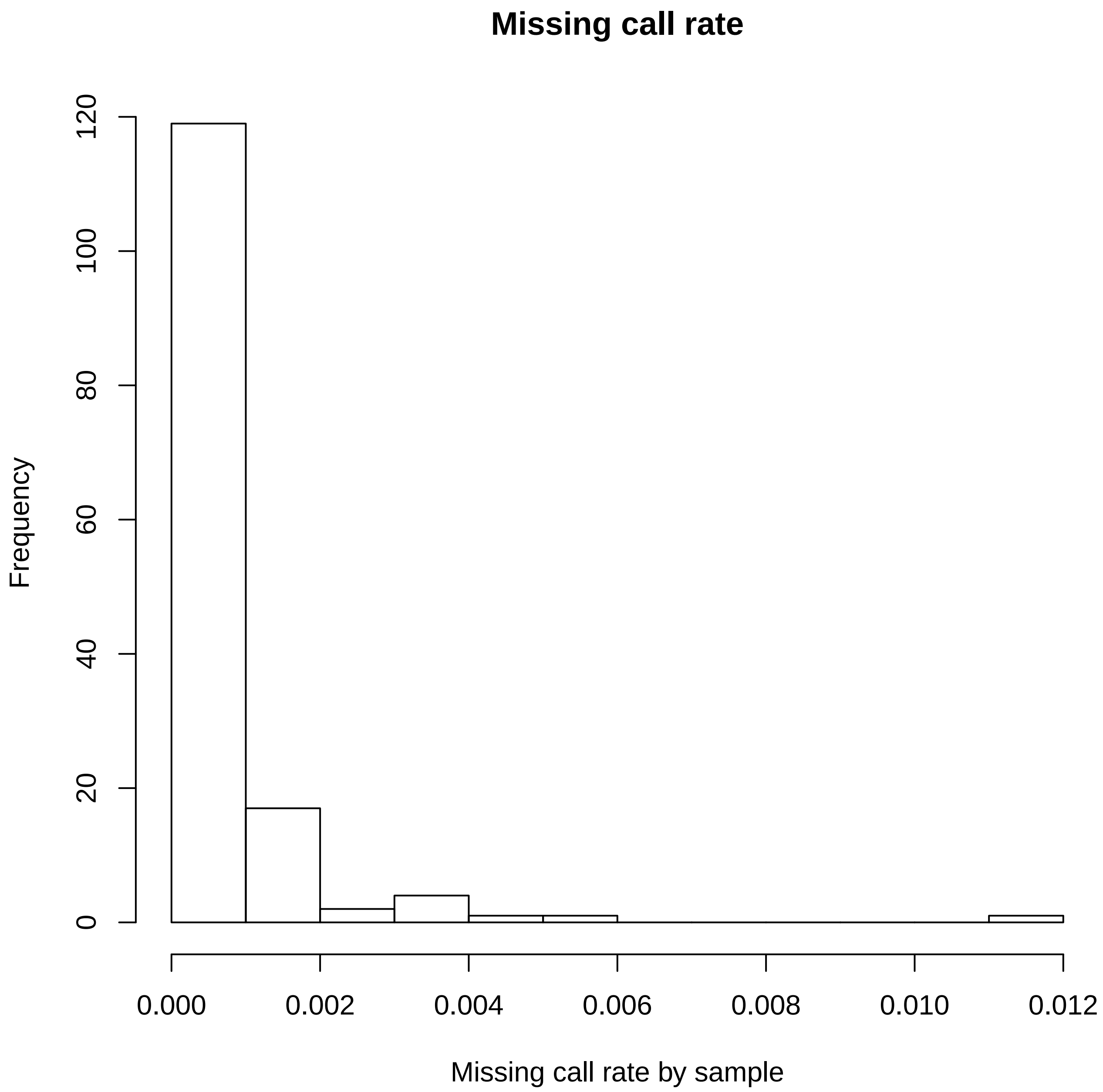


Figure 10: Boxplot of missing call rate for study samples categorized by genotyping plate. Red boxes indicate plates containing samples that failed in the first round of genotyping and were re-genotyped together. The width of each box is proportional to the square root of sample size.

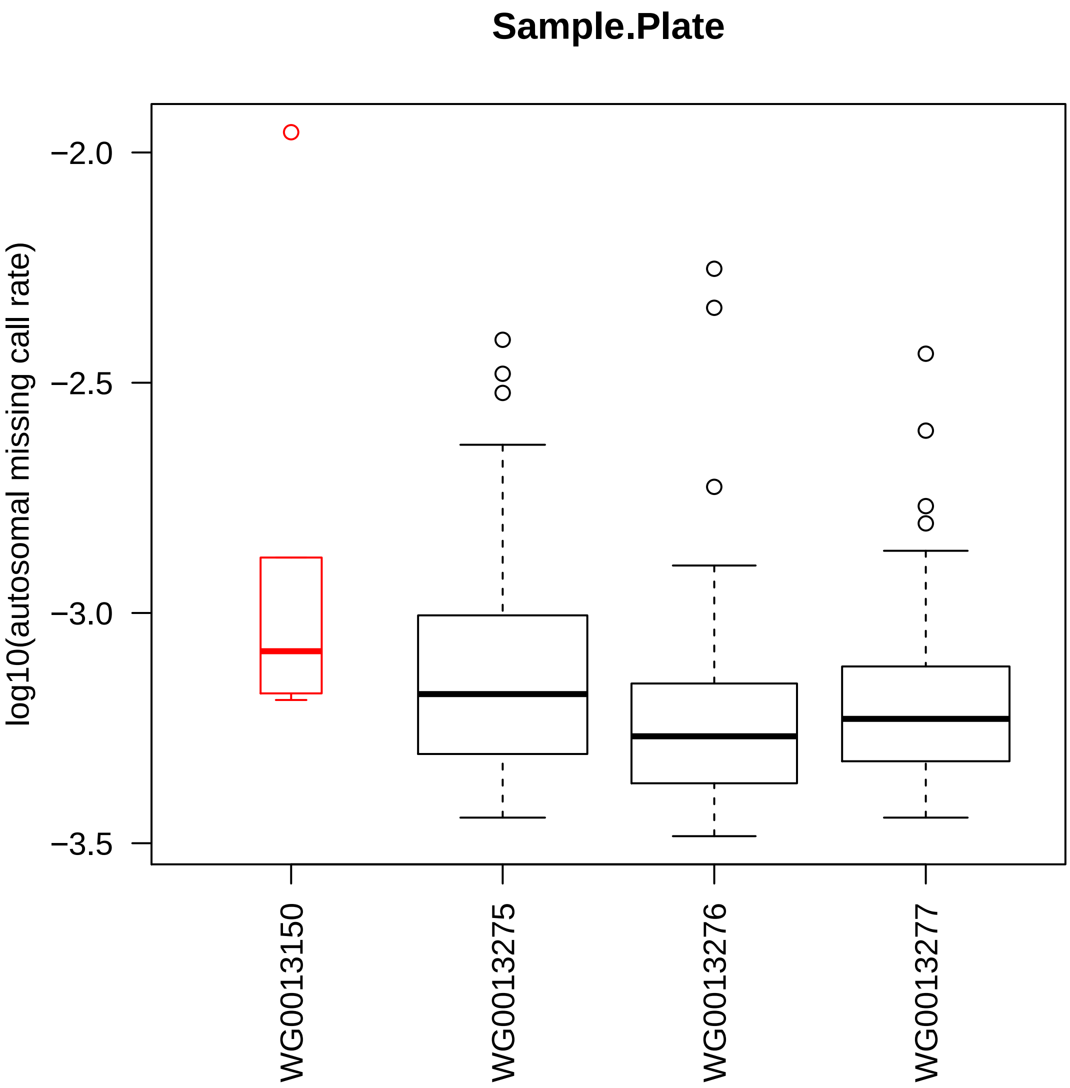
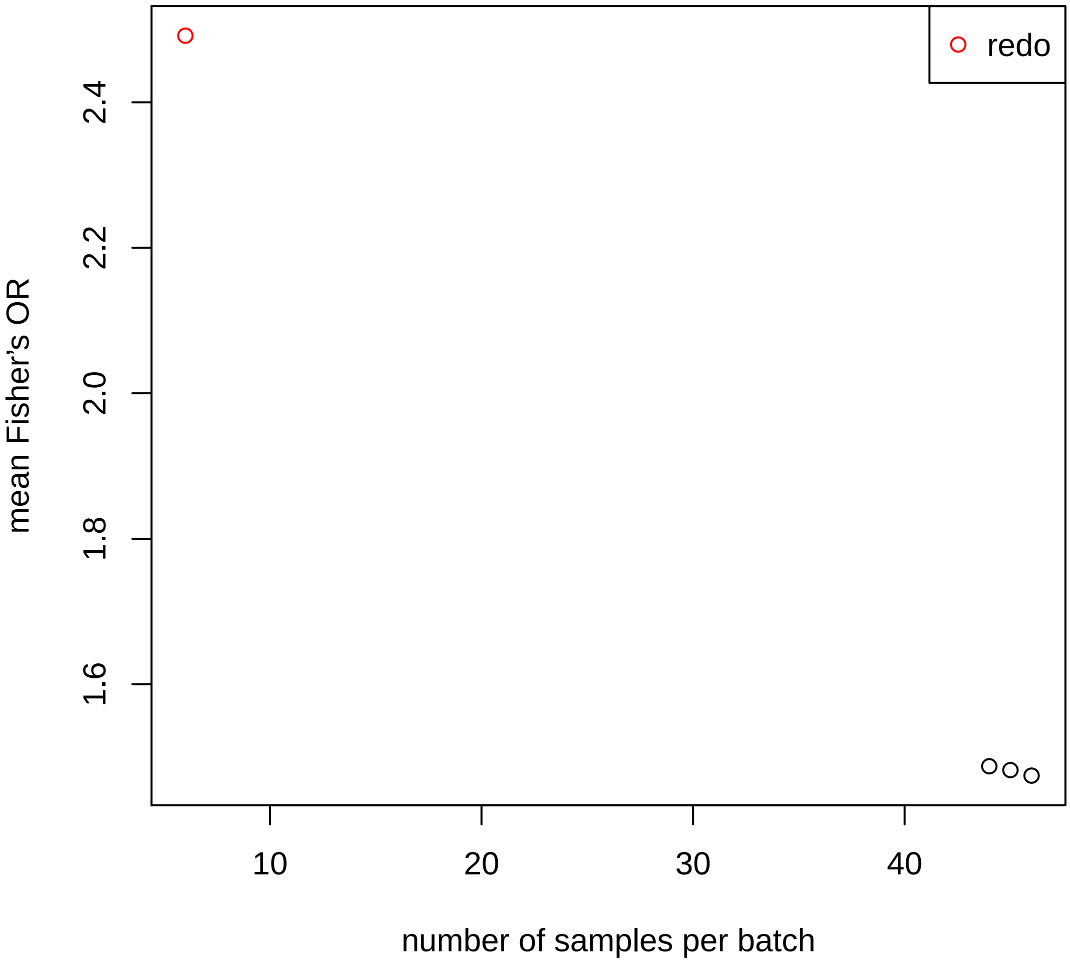
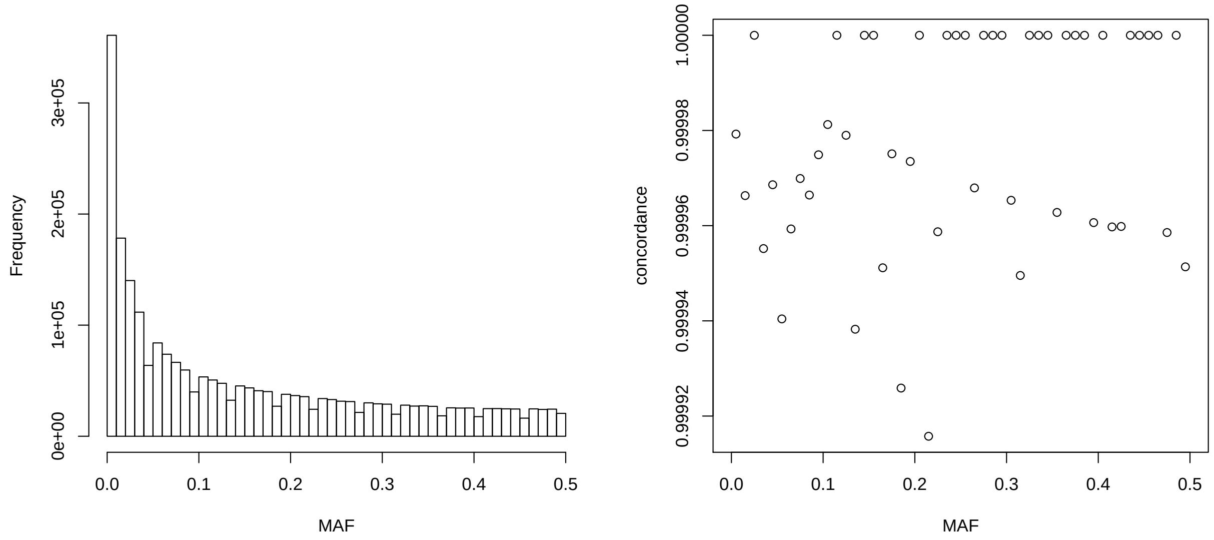


Figure 11: Mean odds ratio (OR) plotted against number of samples per plate. Red points indicate plates containing samples that failed in the first round of genotyping and were re-genotyped together.





(a) Distribution of minor allele frequency.

(b) Overall concordance.

Figure 12: Summary of concordance by SNP over 1 duplicate sample pair, binned by minor allele frequency.

Figure 13: Quantile-quantile plots for $-log_{10}(p)$ from exact test of Hardy-Weinberg equilibrium. 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.

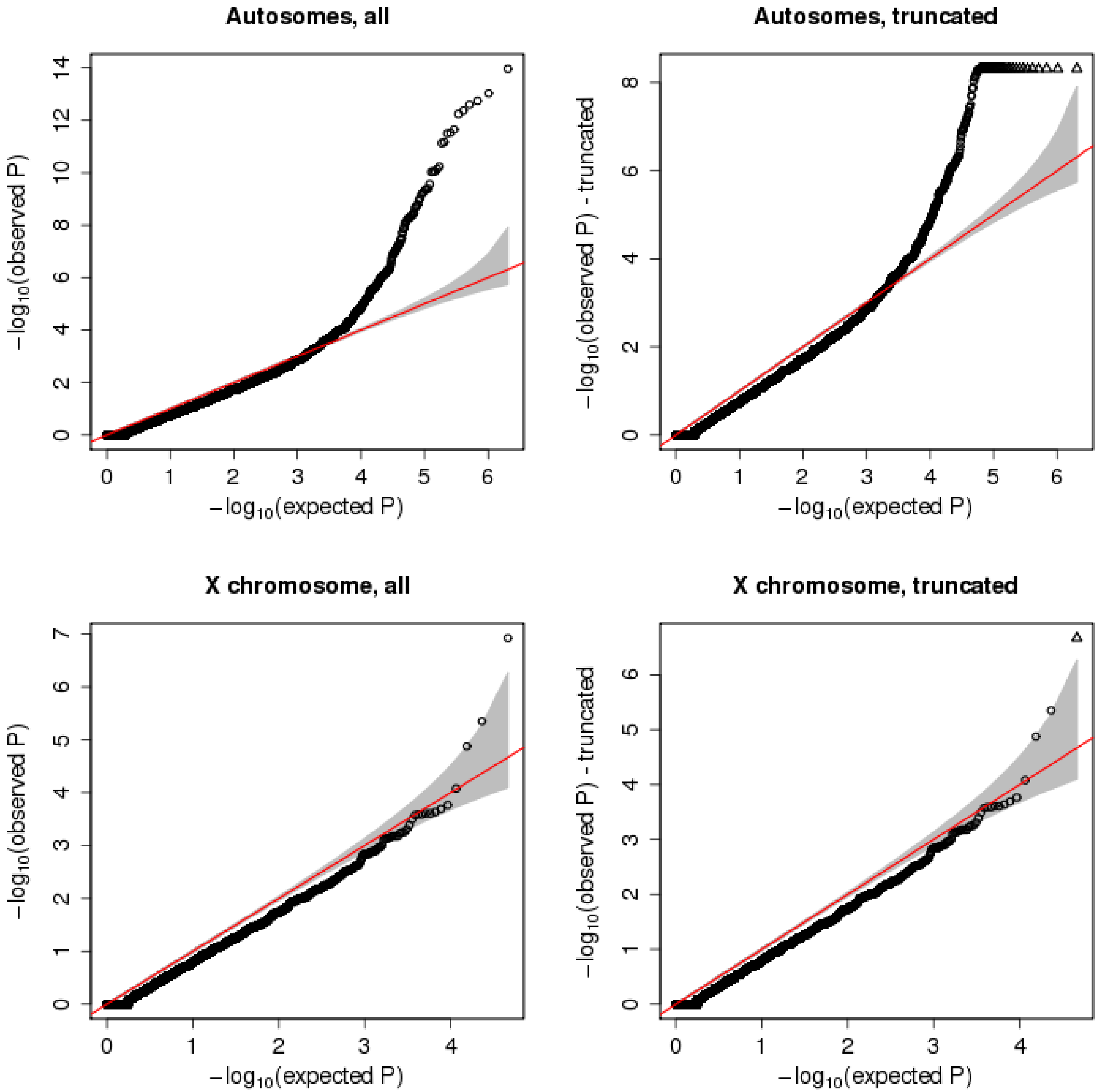


Figure 14: Distributions of estimated inbreeding coefficient for a random sample of 48,654 autosomal SNPs with black representing observed values calculated from the data and red representing values calculated from simulation assuming Hardy-Weinberg equilibrium. The potential values range from -1 to 1.

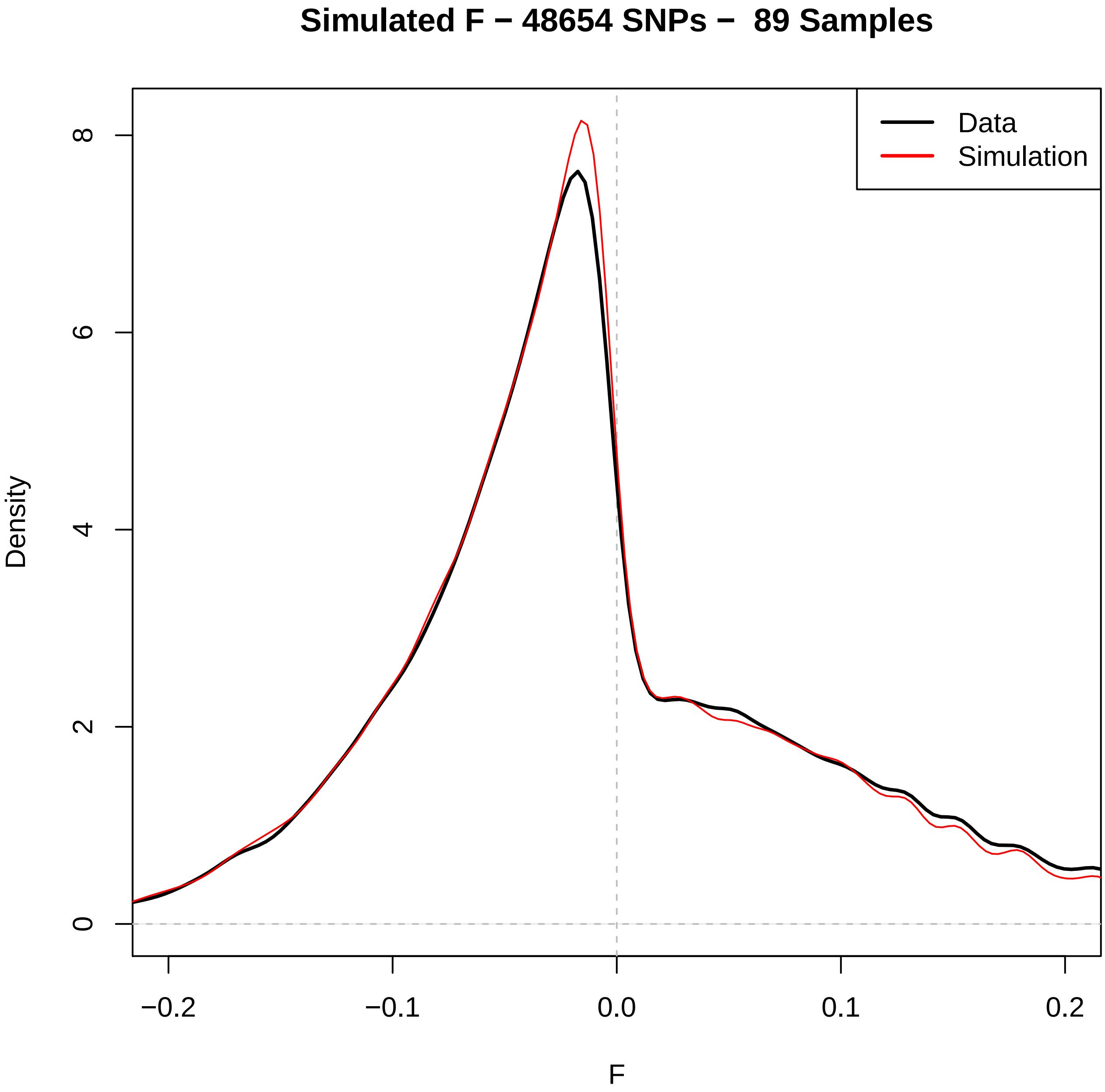
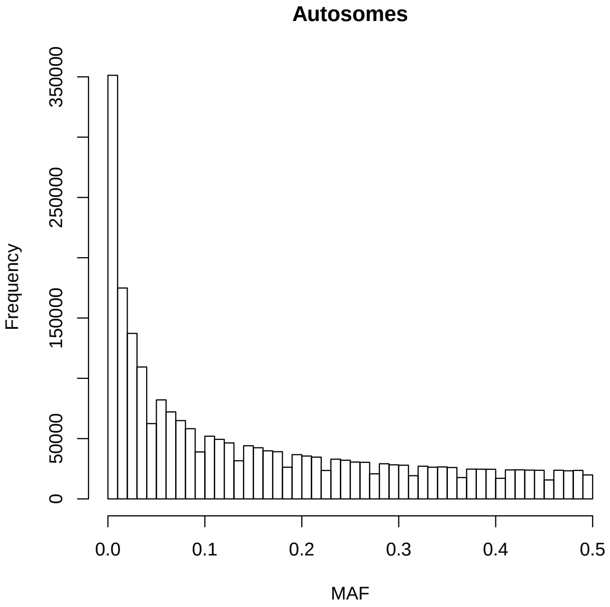
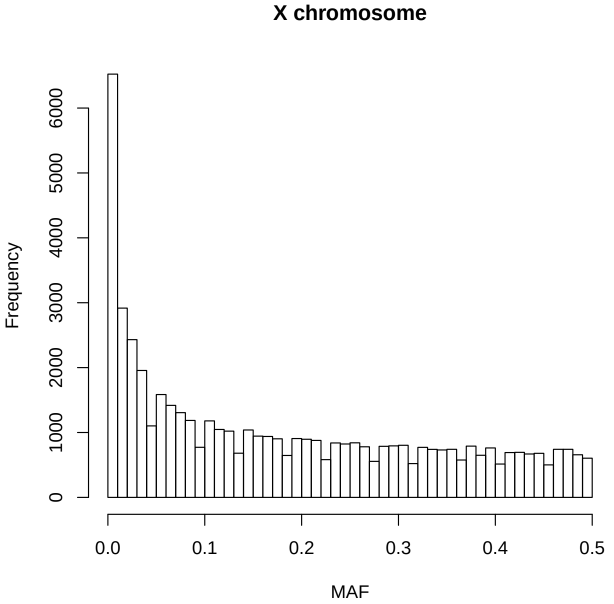


Figure 15: Minor allele frequency distribution across all study subjects.



(a) Autosomes



(b) X chromosome

Figure 16: Quantile-quantile plots for preliminary association tests with trait variable “NA_P.bitter”. QQ plots are provided after using no SNP filter, using composite filter, using composite filter plus MAF filter. The bottom-right plot shows the QQ plot for SNPs whose MAF was less than the MAF filter threshold.

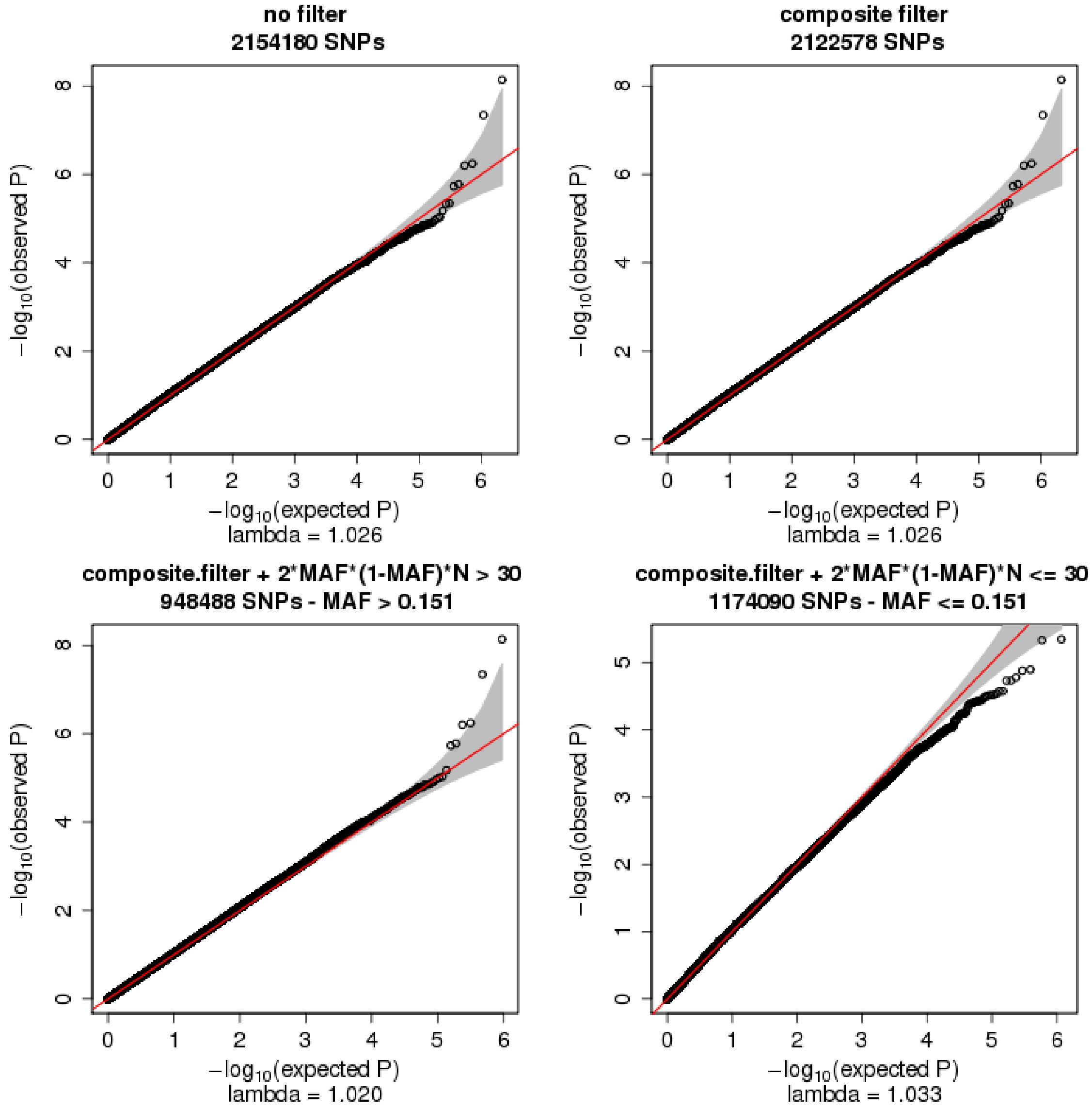
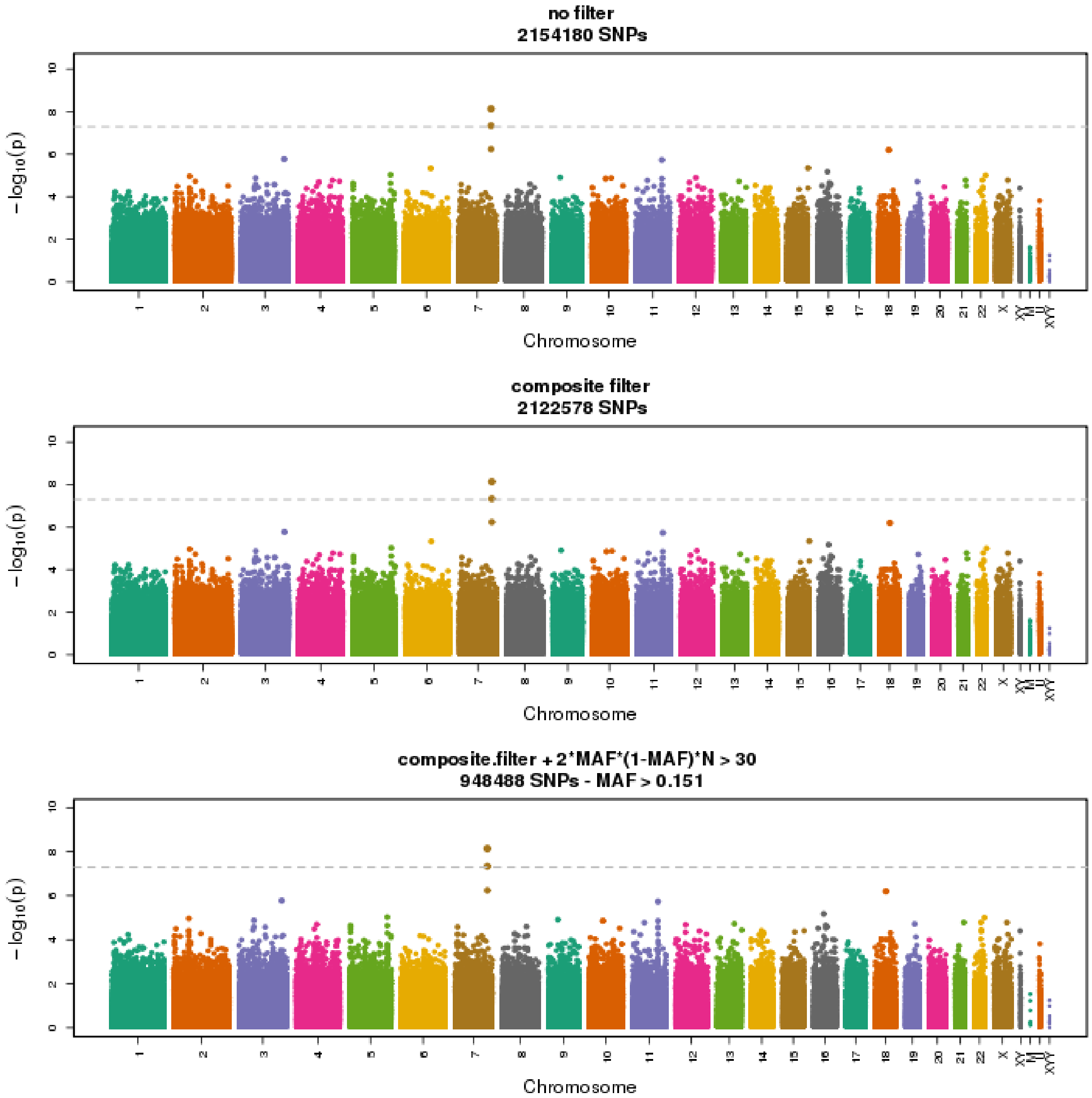


Figure 17: Manhattan plots for preliminary association tests with trait variable “NA_P.bitter”

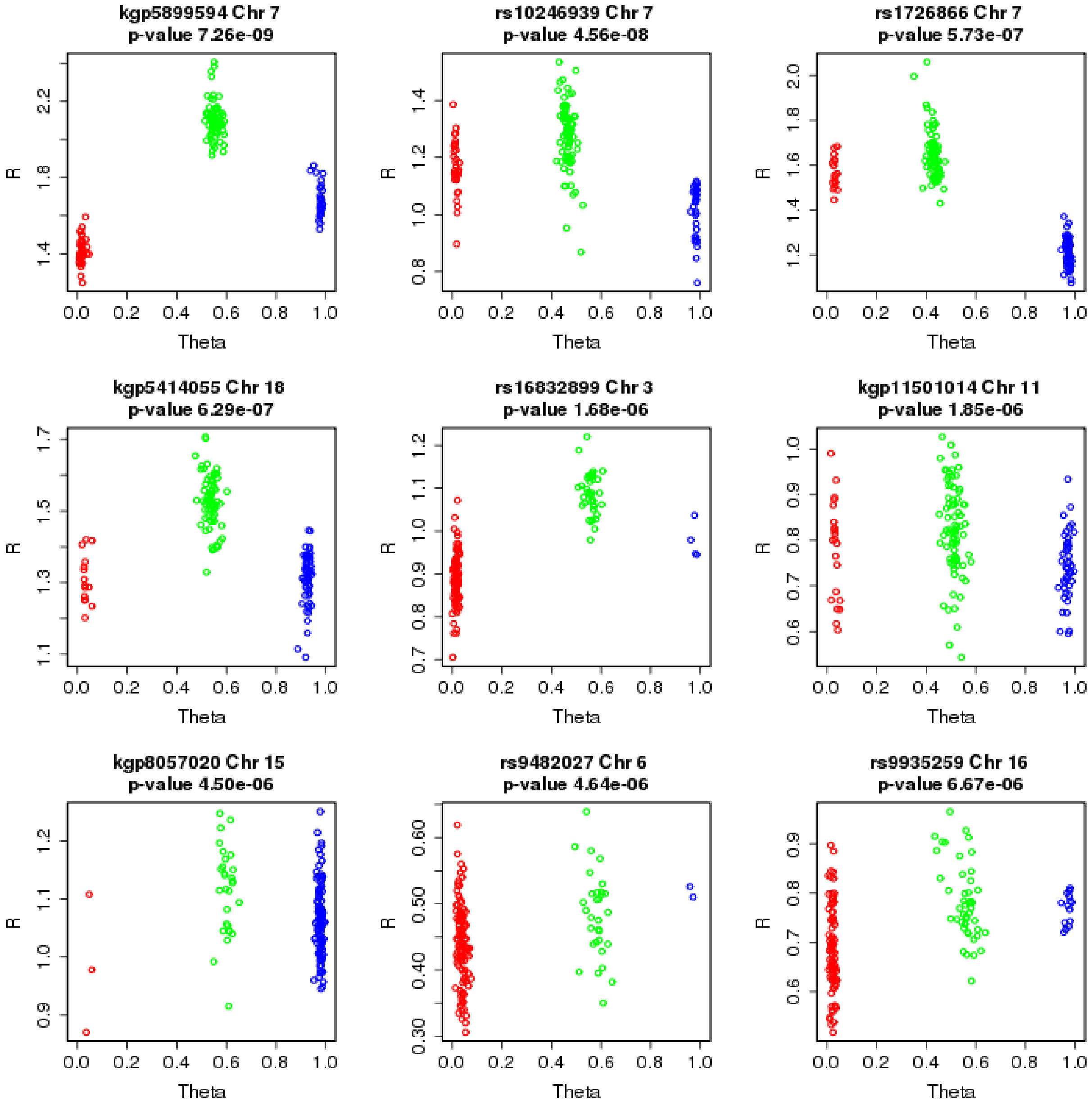


no filter
2154180 SNPs

composite filter
2122578 SNPs

composite.filter + 2*MAF*(1-MAF)*N > 30
948488 SNPs - MAF > 0.151

Figure 18: Genotype cluster plots for the top 9 SNPs from the preliminary association test with trait variable “NA_P.bitter” after applying the composite filter.



kgp5899594 Chr 7
p-value 7.26e-09

rs10246939 Chr 7
p-value 4.56e-08

rs1726866 Chr 7
p-value 5.73e-07

kgp5414055 Chr 18
p-value 6.29e-07

rs16832899 Chr 3
p-value 1.68e-06

kgp11501014 Chr 11
p-value 1.85e-06

kgp8057020 Chr 15
p-value 4.50e-06

rs9482027 Chr 6
p-value 4.64e-06

rs9935259 Chr 16
p-value 6.67e-06

Figure 19: Quantile-quantile plots for preliminary association tests with trait variable "SUC_B.sweet". QQ plots are provided after using no SNP filter, using composite filter, using composite filter plus MAF filter. The bottom-right plot shows the QQ plot for SNPs whose MAF was less than the MAF filter threshold.

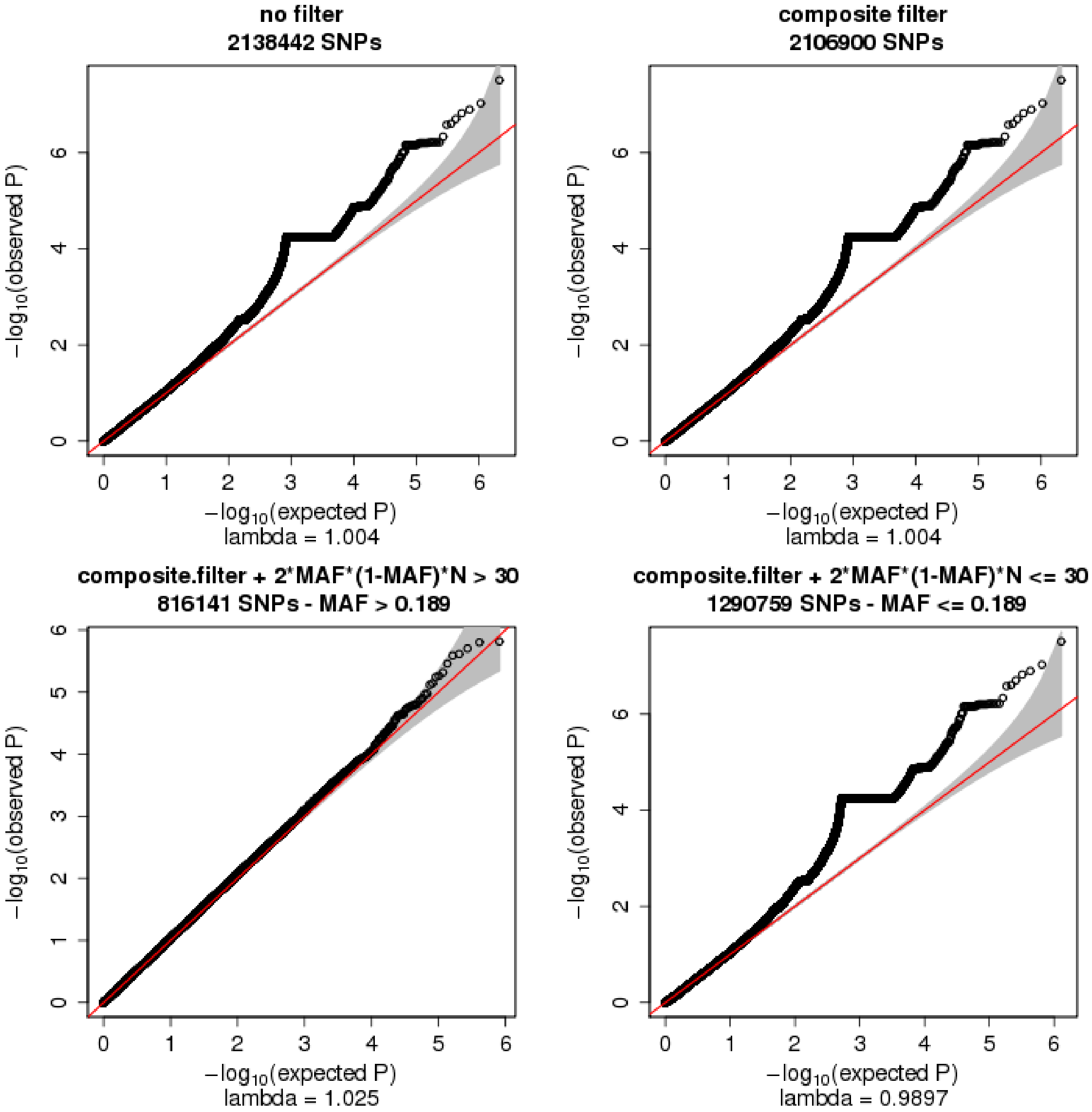
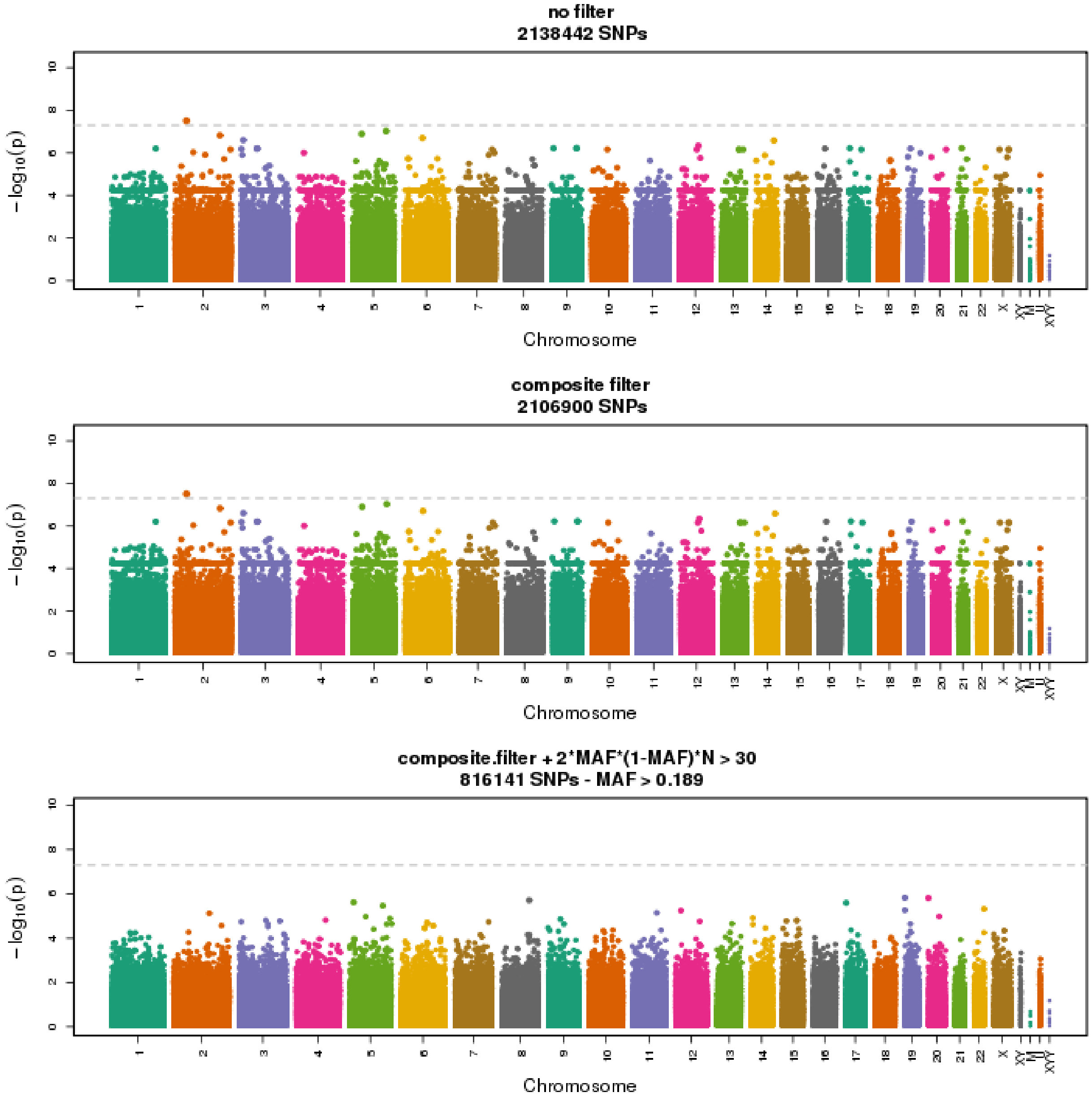


Figure 20: Manhattan plots for preliminary association tests with trait variable “SUC_B.sweet”

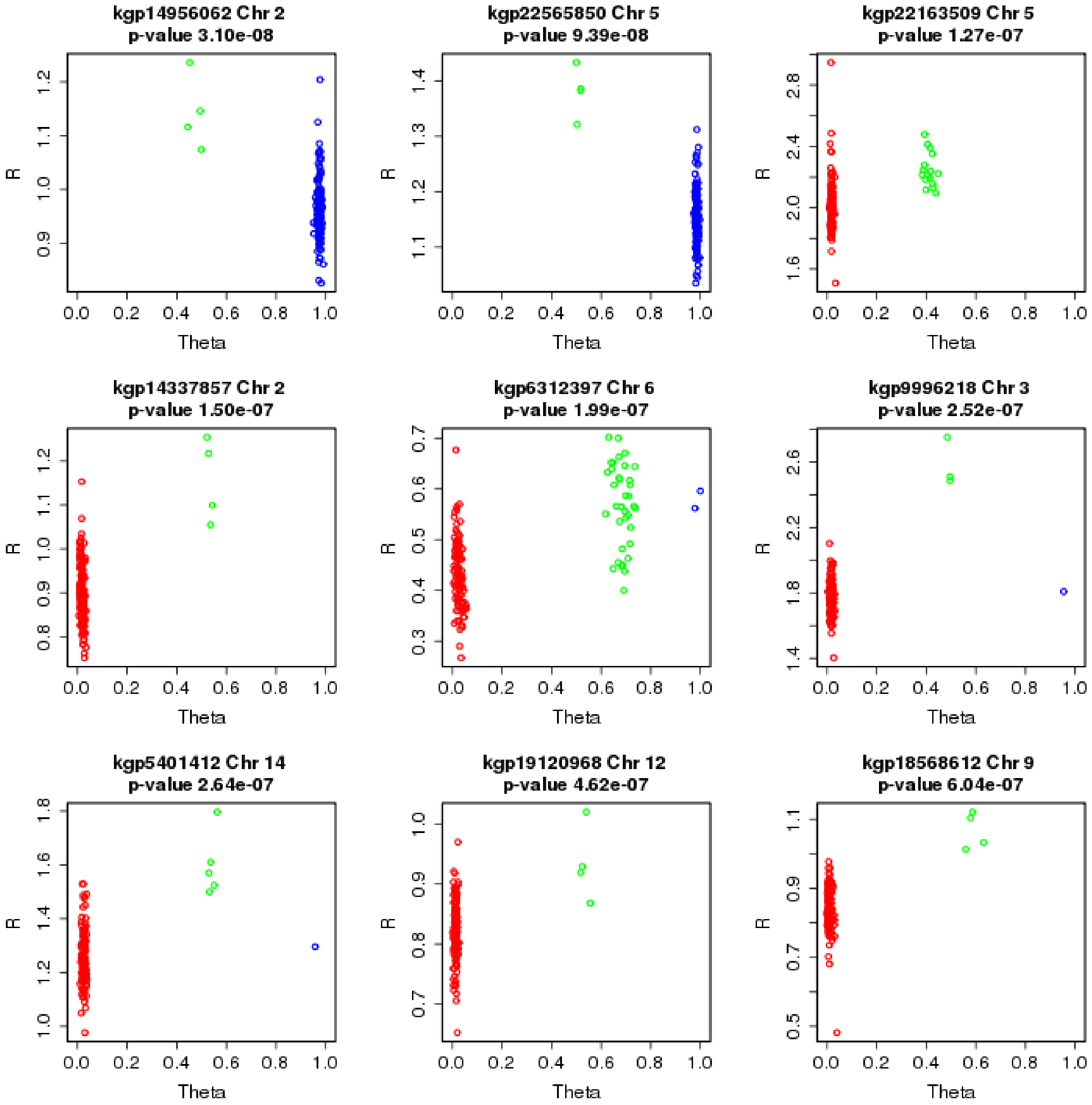


no filter
2138442 SNPs

composite filter
2106900 SNPs

composite.filter + 2*MAF*(1-MAF)*N > 30
816141 SNPs - MAF > 0.189

Figure 21: Genotype cluster plots for the top 9 SNPs from the preliminary association test with trait variable "SUC_B.sweet" after applying the composite filter.


kgp14956062 Chr 2
p-value 3.10e-08
R
Theta
kgp22565850 Chr 5
p-value 9.39e-08
R
Theta
kgp22163509 Chr 5
p-value 1.27e-07
R
Theta
kgp14337857 Chr 2
p-value 1.50e-07
R
Theta
kgp6312397 Chr 6
p-value 1.99e-07
R
Theta
kgp9996218 Chr 3
p-value 2.52e-07
R
Theta
kgp5401412 Chr 14
p-value 2.64e-07
R
Theta
kgp19120968 Chr 12
p-value 4.62e-07
R
Theta
kgp18568612 Chr 9
p-value 6.04e-07
R
Theta