Bipolar Exome Sequencing Pipeline

Duncan Palmer

- Filter genotype content from joint called .vcf. Throw away the following (01_load_vcf_filterGT.py):
  - If homozygous reference, at least one of:
    - Genotype quality < 20
    - Depth < 10
  - If heterozygous, at least one of:
    - (Reference allele depth + alternative allele depth)/depth < 0.8
    - (Alternative allele depth)/depth < 0.2
    - Reference phred-scaled genotype posterior < 20
    - Depth < 10
  - If homozygous variant, at least one of:
    - (Alternative allele depth)/depth < 0.8
    - Reference phred-scaled genotype posterior < 20
    - Depth < 10
- Using 02_prefilter_variants.py, Remove variants that either:
  - Fall in a low complexity region.
  - Fail VQSR.
  - Fall outside of padded target intervals (50bp padding).
  - Filter out invariant sites.
- Using 03_initial_sample_qc.py, run 'sample_qc' using hail and remove:
  - sample call rate < 0.93
  - FREEMIX contamination (%) > 0.02
  - PCT_CHIMERAS (%) > 0.015
  - mean depth (dpMean) < 30

- mean genotype quality (gqMean) < 55

Thresholds used were based on plotting the various distributions. See Figure 1 for CDFs and the thresholds used.

- Export common variants (allele frequency between 0.01 to 0.99) with high call rate (> 0.98) to plink format and prune to independent SNPs using '–indep 50 5 2' (04_export_plink.py).
- Impute the sexes of the individuals (05_impute_sex.py; see Figure 2) with this pruned set of variants on the X, and create list of samples with incorrect or unknown sex as defined by:
  - Sex is 'unknown' in the phenotype files.
  - F-statistic > 0.6 and the sex is 'female' in the phenotype file.
  - F-statistic < 0.6 and the sex is 'male' in the phenotype file.
- Compute IBD between all pairs of individuals using the pruned set of variants in the autosomes (06_ibd.py; see Figure 3).
  - Create sample list of individuals such that no pair has $\hat{\pi} > 0.2$ (06_ibd_filter.r).
- Run PCA on samples after removing relateds and those that passed initial QC, using the pruned set of variants ('09_pca.py'; see Figure 4).
- Run PCA on samples plus all of 1000 genomes (10_pca_1kg.py ; see Figure 6).
  - Train a random forest (10,000 trees) on the super populations of 1000 genomes and predict super populations of BipEx samples.
  - Denote strictly defined European subset as those with probability > 0.95 of being European according to the classifier.
- Run PCA on samples restricted to the strictly defined European subset, and check for outliers (11_pca_EUR.py; see Figure 7).
- Using a much looser definition of European, restrict to US samples from MGH and Johns Hopkins, and run PCA (12_pca_.aj.py; see Figure 8)
  - Identify Ashkenazi Jewish cluster, and create a list of outliers (AJ or otherwise) for downstream removal.
- Run further PCA on:
  1. Strictly defined Europeans and Ashkenazi Jews (13_pca_aj_1kg.py; see Figure 9).
     - Use Ashkenazi Jewish cluster to train a random forest and determine if there are further Ashkenazi Jews in the remainder of the dataset - there weren't.
  2. Strictly defined Europeans (13_pca_EUR_1kg.py; see Figure 10)

- Now restrict to samples in the strictly defined European subset, filter to the unrelated list, and filter out samples with incorrectly defined sex or unknown sex, and run variant QC (14_final_variant_qc.py ; see Figure 11). Remove variants that satisfy at least one of:
  - Invariant site in cleaned sample subset.
  - Call rate < 0.97
  - Control call rate < 0.97
  - Case call rate < 0.97
  - |Case call rate - Control call rate| > 0.02
  - p-value for Hardy Weinberg Equilibrium $< 10^{-6}$
- Remove sample outliers after the variant cleaning in the previous step (see Figure 12 for their distributions). Remove sample if at least one of
  - Ratio of heterozygous to homozygous variant
  - Ratio of insertions to deletions
  - Ratio of transitions to transversions

  lies more than three standard deviations away from the mean.
- Export common (0.01 - 0.99) variants to plink format, prune and evaluate final principal components for downstream analysis, and save cleaned .mt files.
- Run final PCA, pruning the cleaned set of variants and cleaned set of samples (see Figure 13).

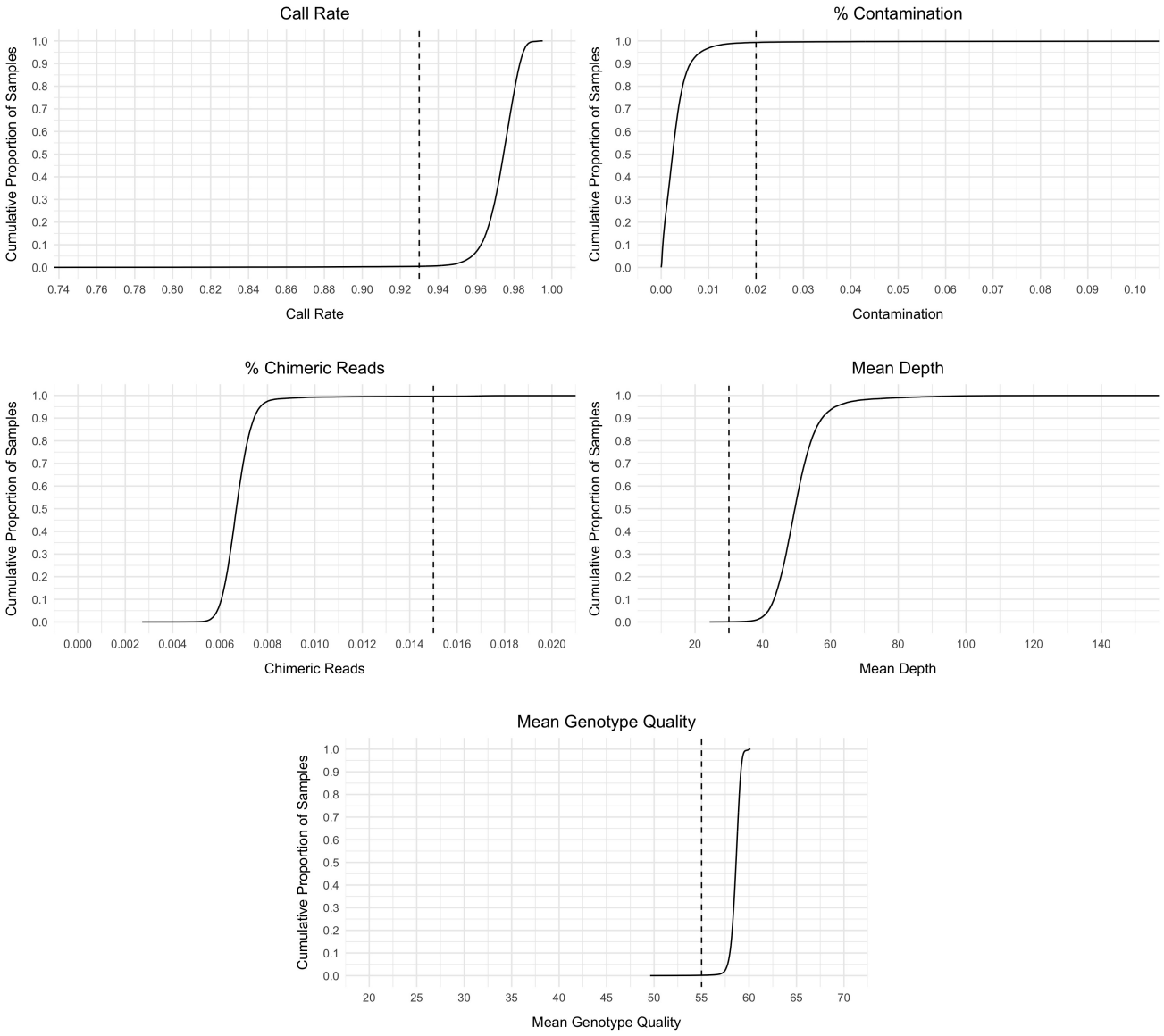


Figure 1: CDFs of various metrics of sample quality with thresholds for inclusion/exclusion.

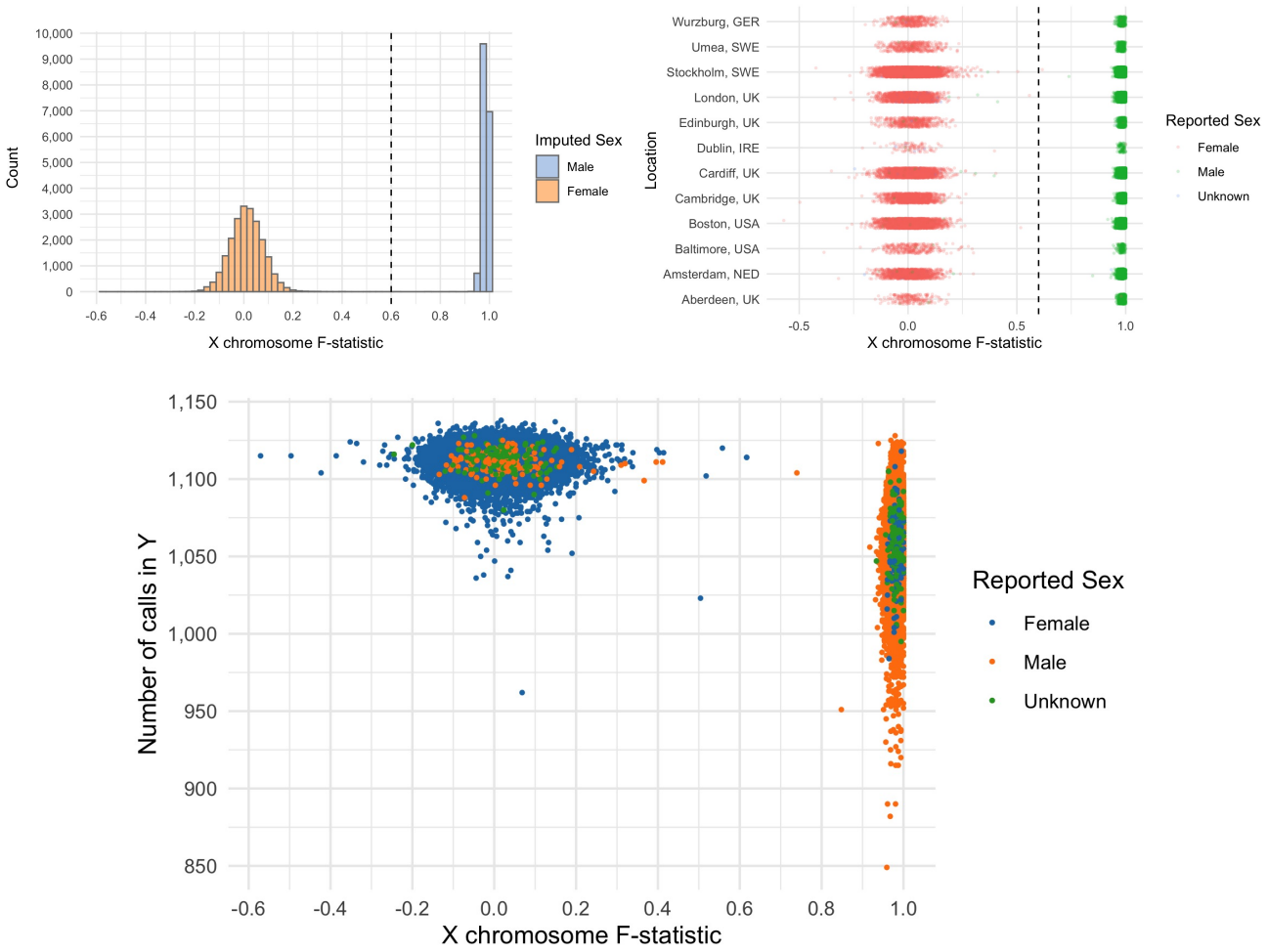


Figure 2: F-statistic used to determine if individual is male or female. Threshold set at 0.6.

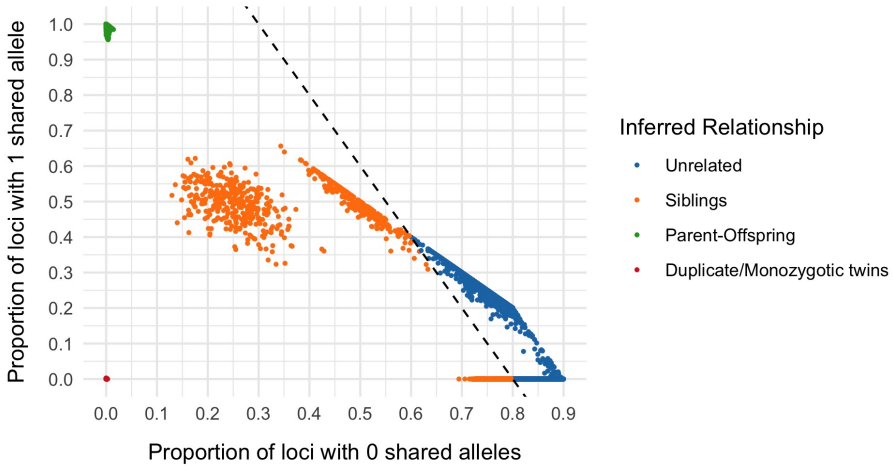


Figure 3: IBD plot. Threshold for 'related' set at $\hat{\pi} > 0.2$.

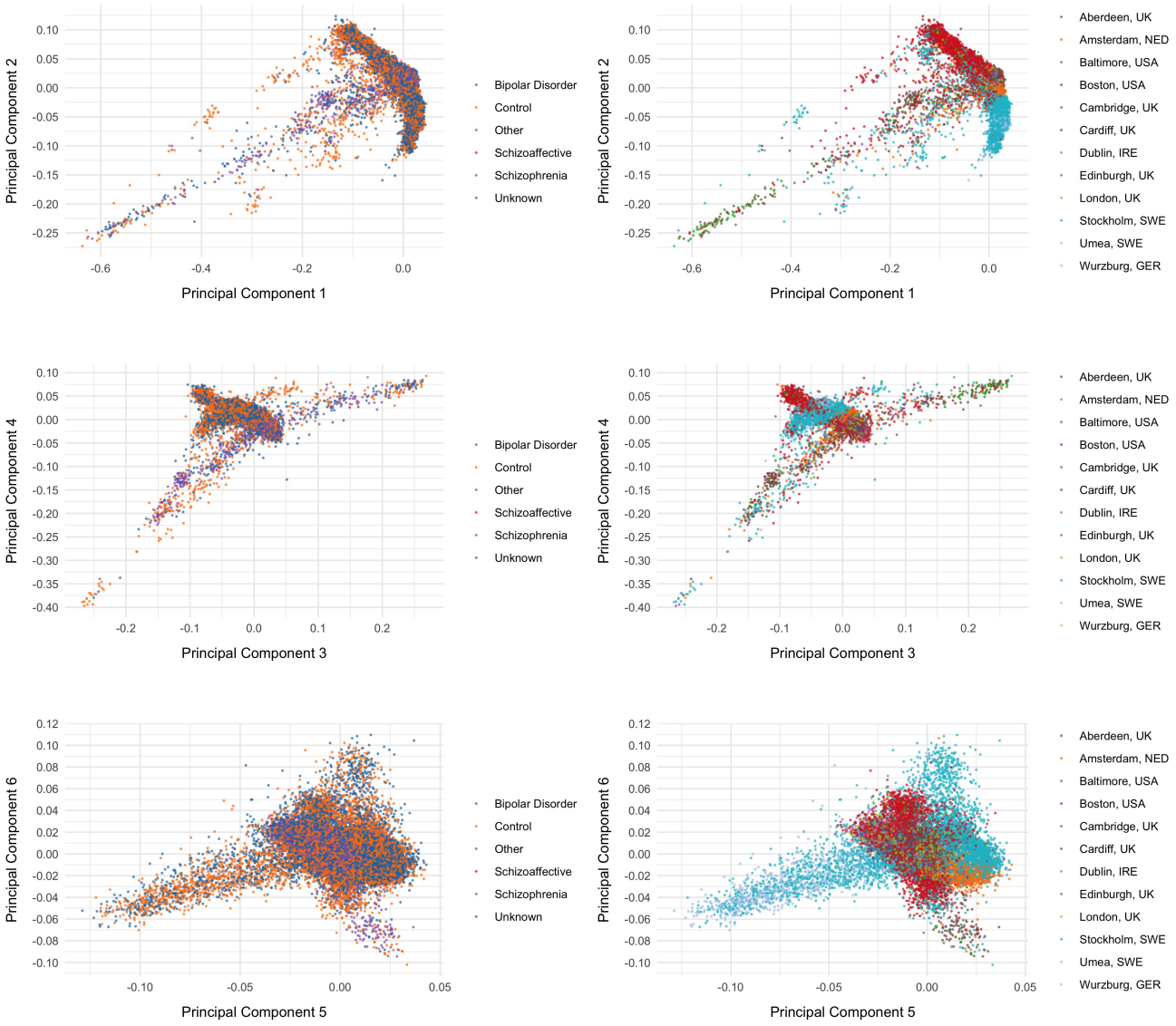


Figure 4: Initial PCA of all samples

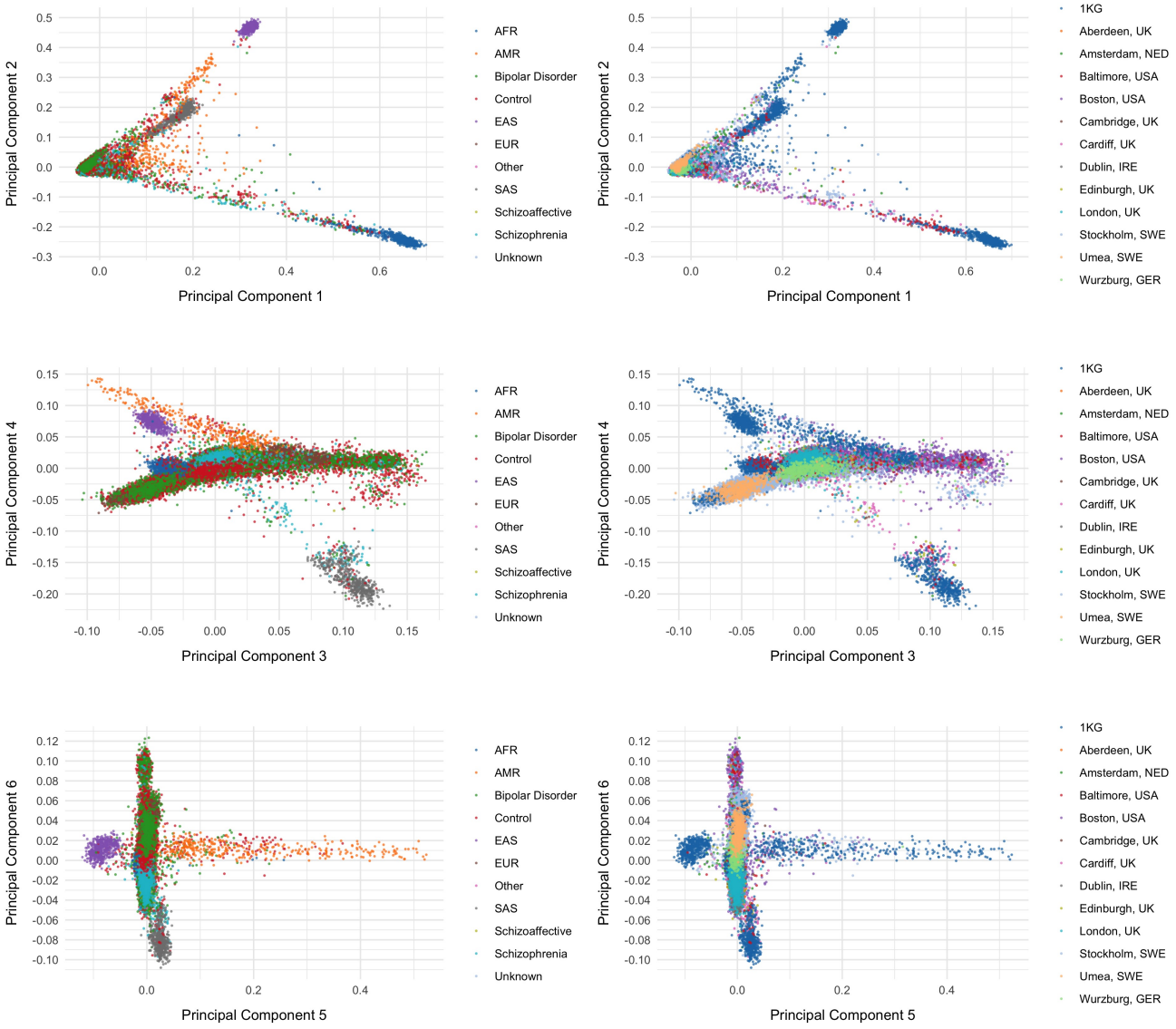


Figure 5: Initial PCA of all samples with 1000 Genomes

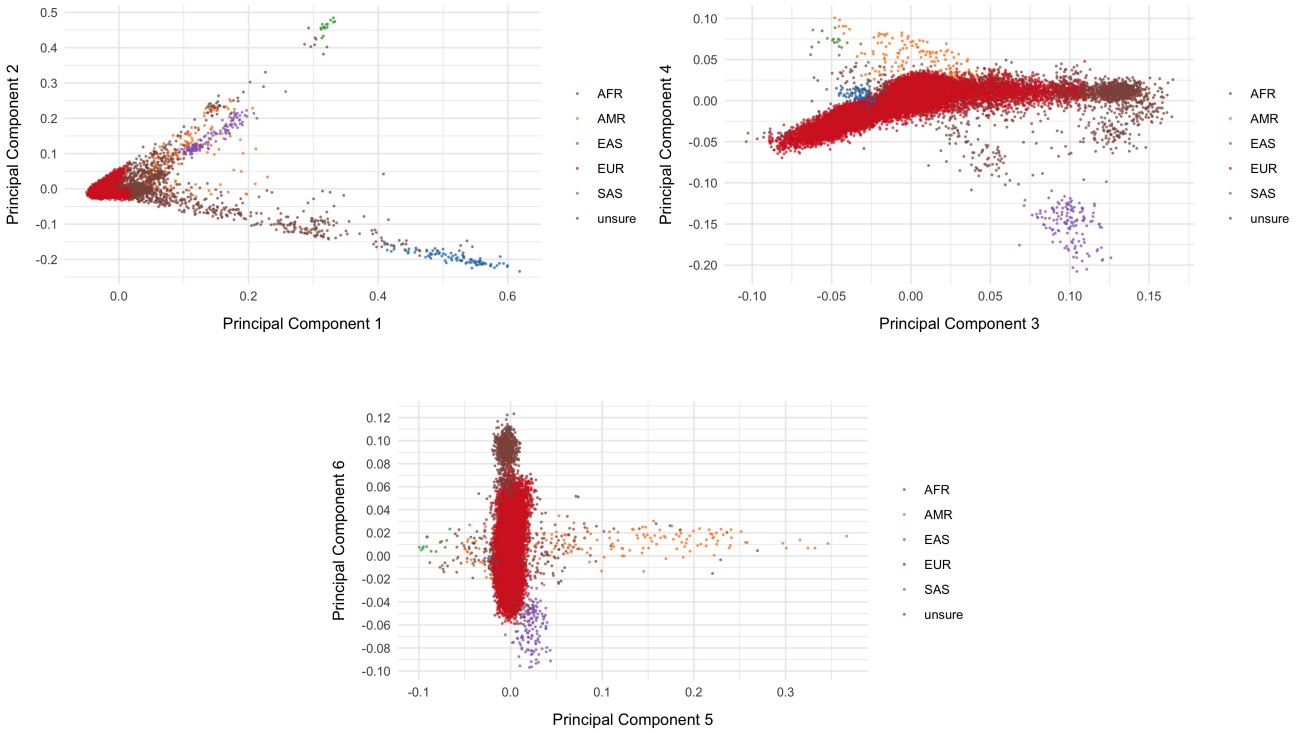


Figure 6: Strict definition of European samples using random forest classifier, 500 trees. $P(European) > 0.95$.

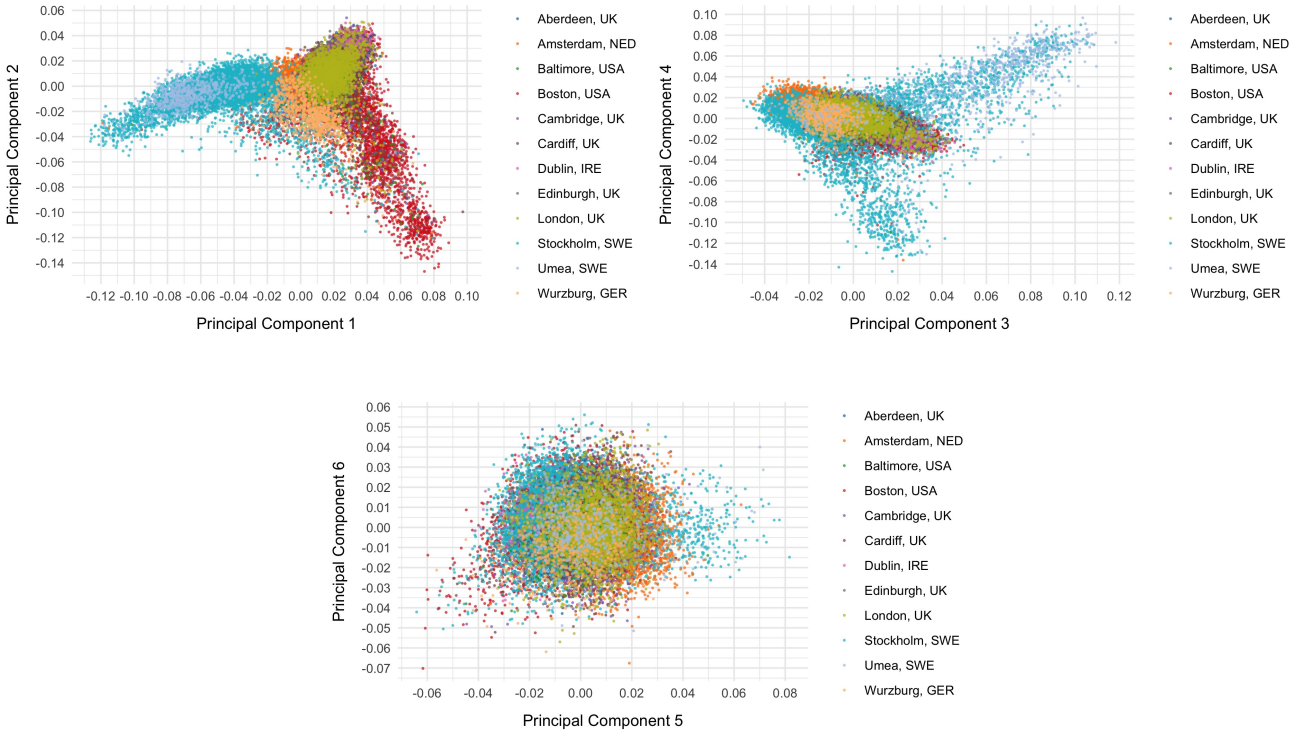
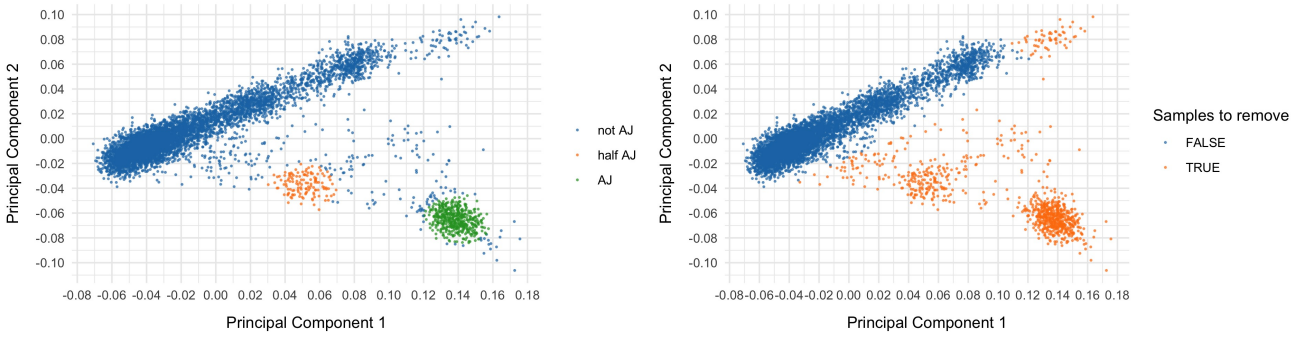


Figure 7: PCA after restricting to strict European subset.



(a) AJ and half AJ cluster identified

(b) Removal of outliers in US European samples

Figure 8: PCA of loosely defined 'European' US samples to identify AJ cluster

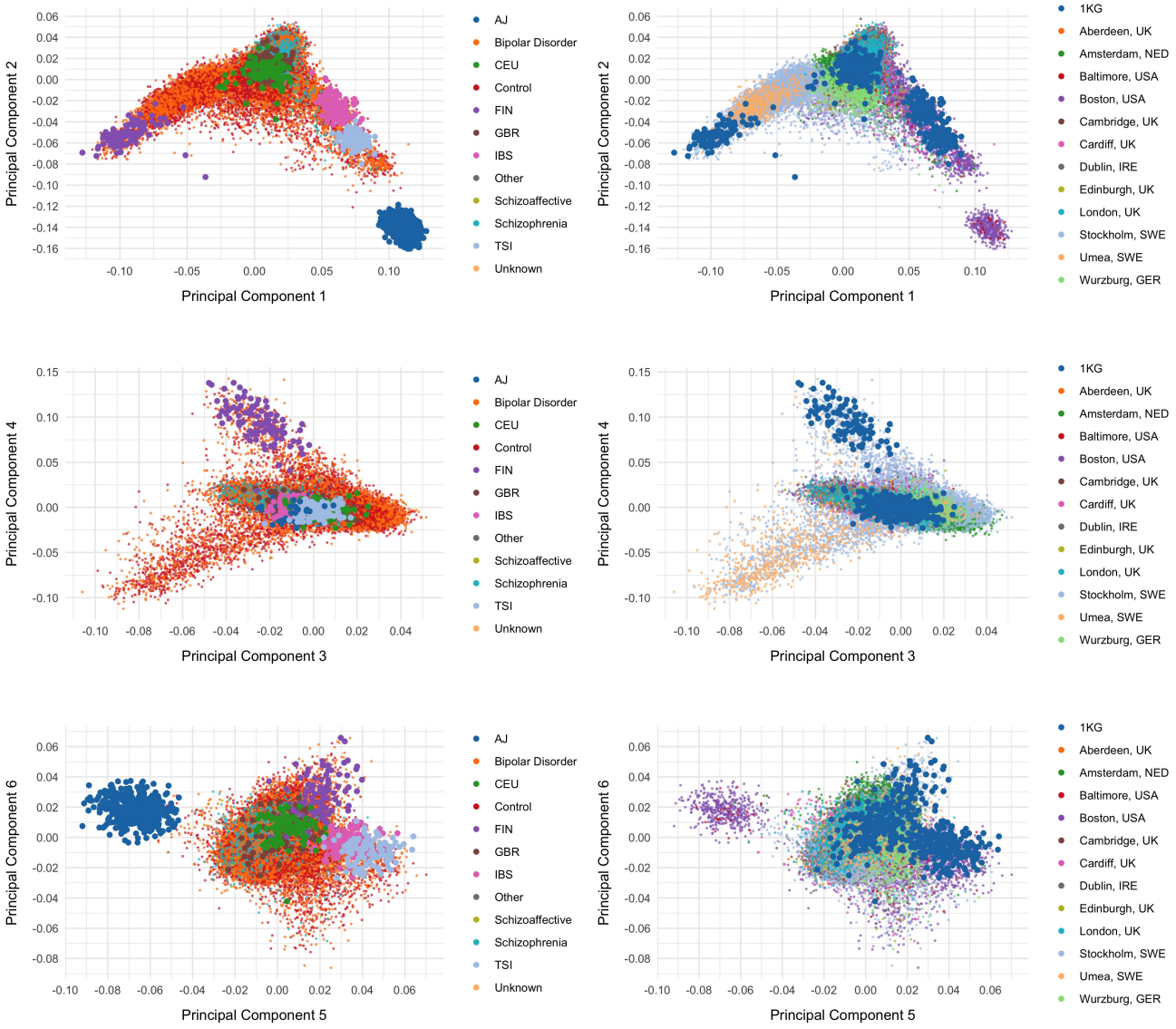


Figure 9: PCA of European cohorts with 1000 Genomes

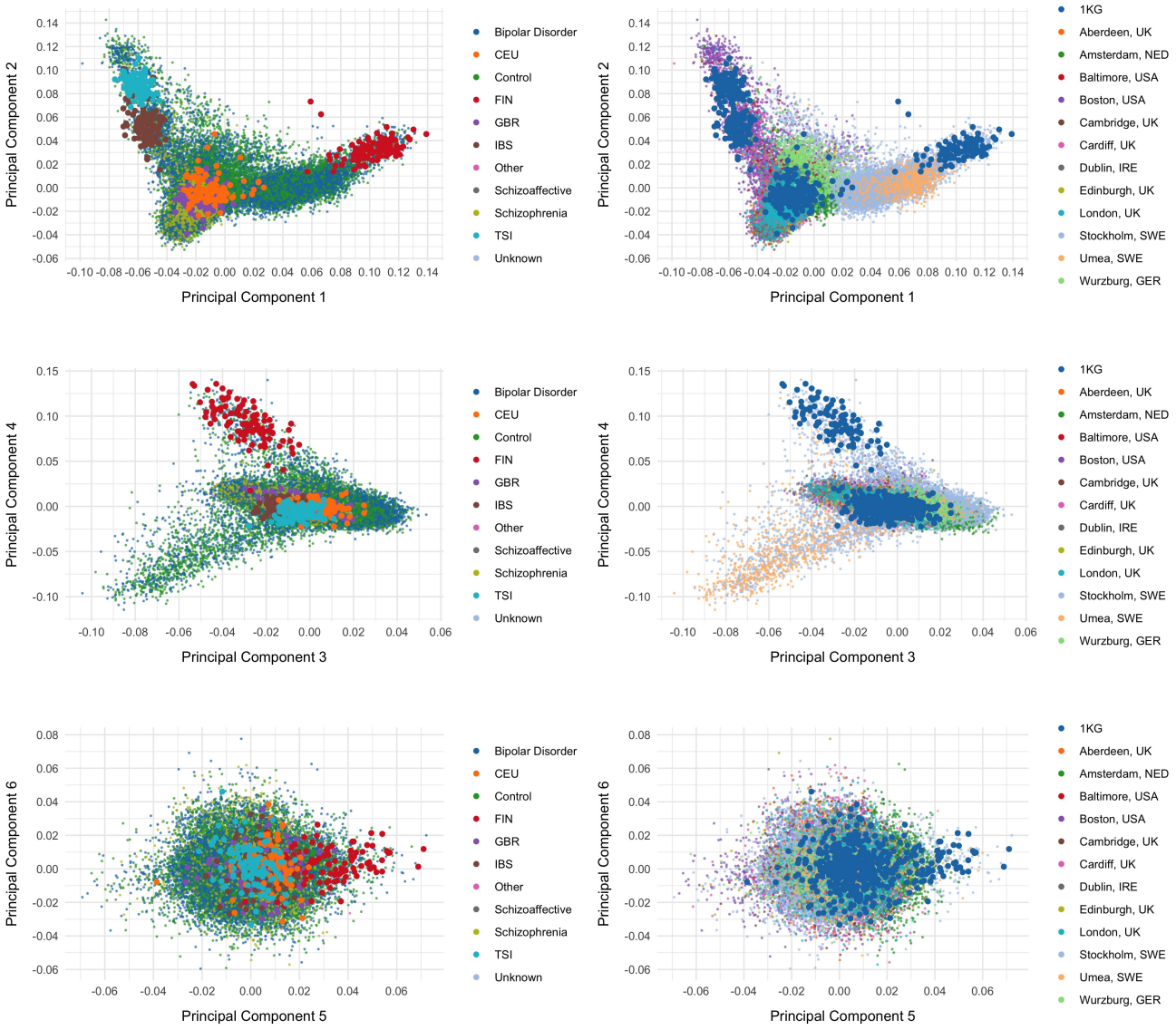


Figure 10: PCA of European cohorts with 1000 Genomes

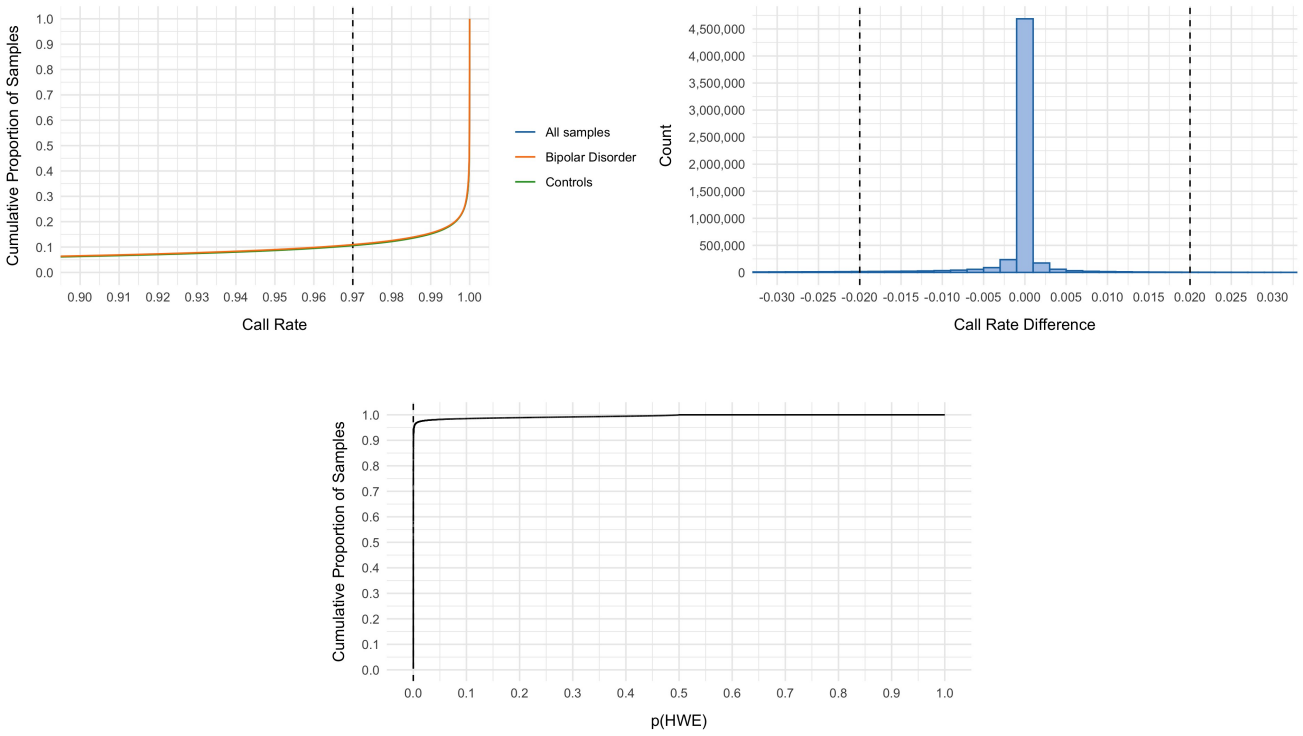


Figure 11: Variant QC

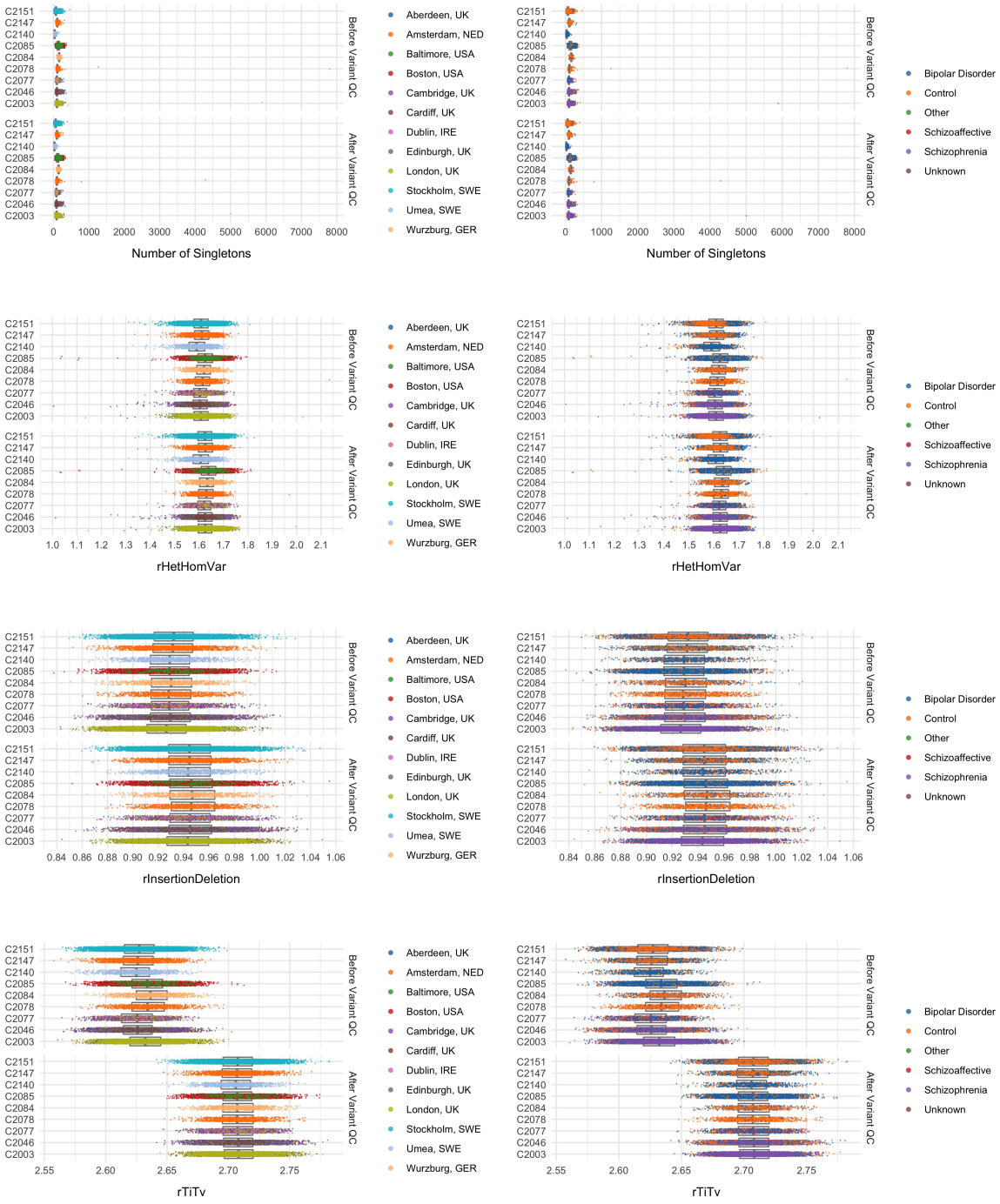


Figure 12: Visualise the impact of variant QC on different metrics.

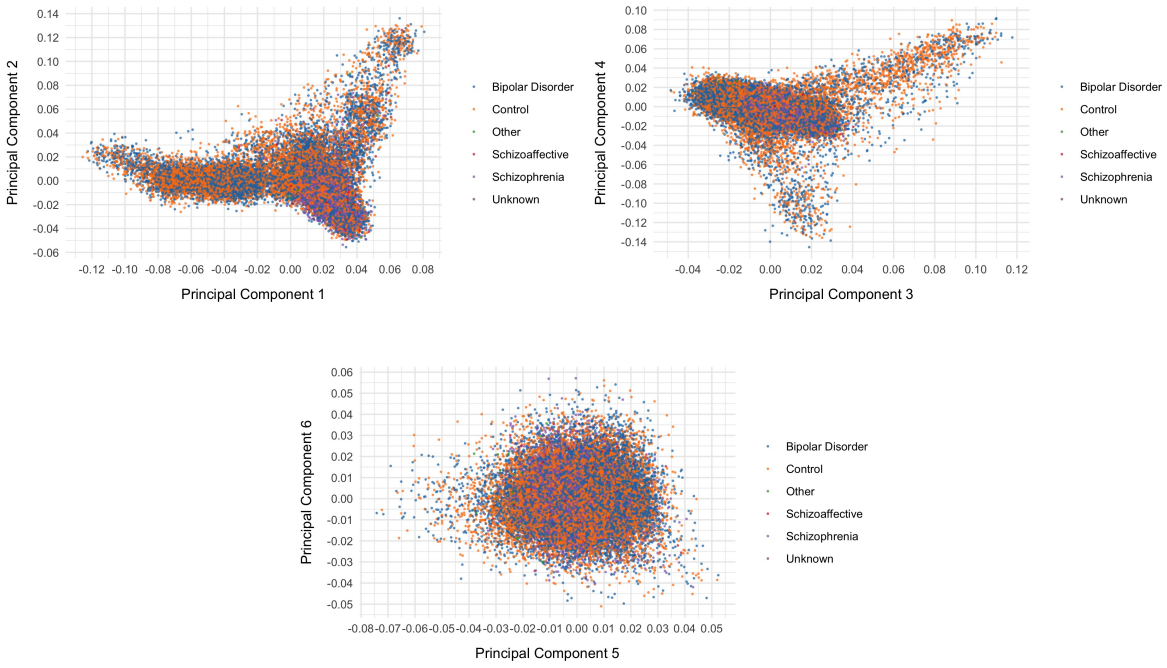


Figure 13: PCA of pruned cleaned variants and samples.