AMP-DCC Quality Control Report

NUS

10/09/2019 (03:53)

Prepared by Ryan Koesterer on behalf of the AMP-DCC Data Analysis Team at Broad Institute

Contact: amp-dcc-dat@broadinstitute.org

This document was generated using Loamstream [10] and the AMP-DCC Data Analysis Pipeline [11]

Contents

1 Introduction

This document contains details of our in-house quality control procedure and its application to the NUS dataset. We received genotypes for 13,372 unique samples distributed across 7 different genotype arrays. Quality control was performed on these data to detect samples and variants that did not fit our standards for inclusion in association testing. After harmonizing with modern reference data, the highest quality variants were used in a battery of tests to assess the quality of each sample. Duplicate pairs, samples exhibiting excessive sharing of identity by descent, samples whose genotypic sex did not match their clinical sex, and outliers detected among several sample-by-variant statistics have been flagged for removal from further analysis. Additionally, genotypic ancestry was inferred with respect to a modern reference panel, allowing for variant filtering and association analyses to be performed within population as needed.With the exception of inferring each samples ancestry, QC was performed on each array separately as much as possible, allowing for flexibility in the way the data can be used in downstream analyses.

2 Data

2.1 Samples

Initially, the array was checked for sample genotype missingness. Any samples with extreme genotype missingness (> 0.5) were removed prior to our standard quality control procedures. There were no samples removed from this data set.

The following diagram (Figure 1) describes the remaining sample distribution over the 7 genotype arrays, along with their intersection sizes.

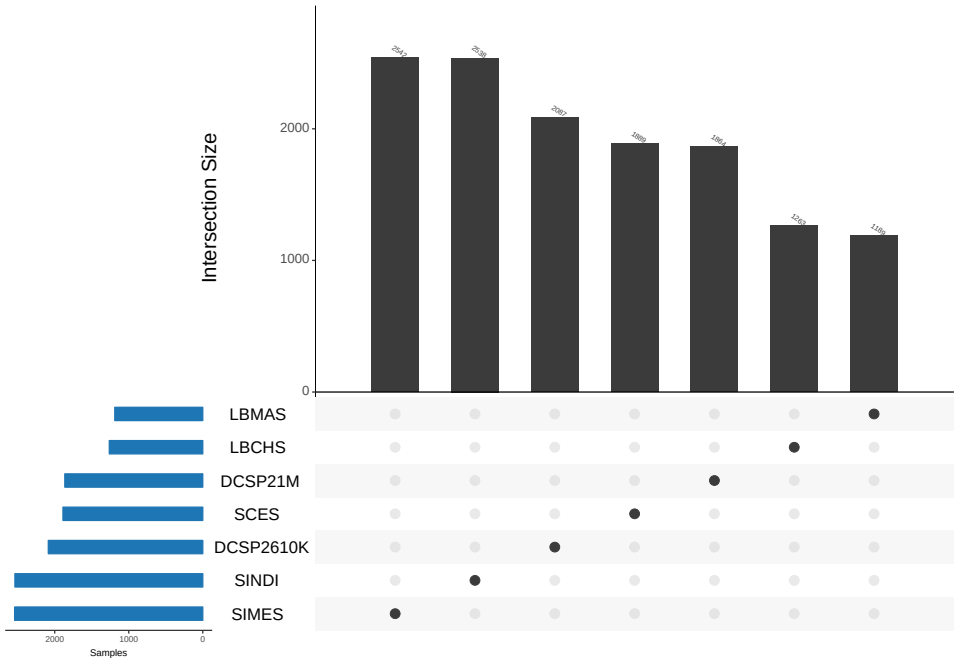


Figure 1: Samples distributed by genotyping array

2.2 Variants

Table 1 gives an overview of the different variant classes and how they distributed across allele frequencies for each dataset. Note that the totals reflect the sum of the chromosomes only. A legend has been provided below the table for further inspection of the class definitions.

Table 1: Summary of raw variants by frequency and classification

Freq = Minor allele frequency (MAF) range
Unpl = Chromosome = 0
Auto = Autosomal variants
X = X chromosome non-pseudoautosomal region (non-PAR) variants
Y = Y chromosome variants
X(PAR) = X chromosome pseudoautosomal (PAR) region variants
Mito = Mitochondrial variants
InDel = Insertion/Deletion variants (I/D or D/I alleles)
Multi = Multiallelic variants (2 or more alternate alleles)
Dup = Duplicated variants with respect to position and alleles

	Freq	Unpl	Auto	X	Y	X(PAR)	Mito	InDel	Multi	Dup	Total
DCSP21M	[0]	0	185	0	0	0	0	0	0	0	185
	(0,0.01)	0	98408	0	0	0	0	0	0	0	98408
	[0.01,0.03)	0	45966	0	0	0	0	0	0	0	45966
	[0.03,0.05)	0	40475	0	0	0	0	0	0	0	40475
	[0.05,0.10)	0	94202	0	0	0	0	0	0	0	94202
	[0.10,0.50]	0	650162	0	0	0	0	0	0	0	650162
	Total	0	929398	0	0	0	0	0	0	0	929398
DCSP2610K	[0]	0	635	0	0	0	0	0	0	0	635
	(0,0.01)	0	47853	0	0	0	0	0	0	0	47853
	[0.01,0.03)	0	24003	0	0	0	0	0	0	0	24003
	[0.03,0.05)	0	21478	0	0	0	0	0	0	0	21478
	[0.05,0.10)	0	53134	0	0	0	0	0	0	0	53134
	[0.10,0.50]	0	388095	0	0	0	0	0	0	0	388095
	Total	0	535198	0	0	0	0	0	0	0	535198
LBCHS	[0]	0	61152	0	0	0	0	0	0	0	61152
	(0,0.01)	0	47859	0	0	0	0	0	0	0	47859
	[0.01,0.03)	0	26970	0	0	0	0	0	0	0	26970
	[0.03,0.05)	0	25607	0	0	0	0	0	0	0	25607
	[0.05,0.10)	0	65020	0	0	0	0	0	0	0	65020
	[0.10,0.50]	0	457283	0	0	0	0	0	0	0	457283
	Total	0	683891	0	0	0	0	0	0	0	683891
LBMAS	[0]	0	34812	0	0	0	0	0	0	0	34812
	(0,0.01)	0	45826	0	0	0	0	0	0	0	45826
	[0.01,0.03)	0	36276	0	0	0	0	0	0	0	36276
	[0.03,0.05)	0	27960	0	0	0	0	0	0	0	27960
	[0.05,0.10)	0	66835	0	0	0	0	0	0	0	66835
	[0.10,0.50]	0	472182	0	0	0	0	0	0	0	472182
	Total	0	683891	0	0	0	0	0	0	0	683891

Continued on next page ...

Table 1: ... Continued from previous page

	Freq	Unpl	Auto	X	Y	X(PAR)	Mito	InDel	Multi	Dup	Total
SCES	[0]	0	0	0	0	0	0	0	0	0	0
	(0,0.01)	0	46169	0	0	0	0	0	0	0	46169
	[0.01,0.03)	0	24007	0	0	0	0	0	0	0	24007
	[0.03,0.05)	0	21321	0	0	0	0	0	0	0	21321
	[0.05,0.10)	0	53009	0	0	0	0	0	0	0	53009
	[0.10,0.50]	0	387610	0	0	0	0	0	0	0	387610
	Total	0	532116	0	0	0	0	0	0	0	532116
SIMES	[0]	0	0	0	0	0	0	0	0	0	0
	(0,0.01)	0	42083	0	0	0	0	0	0	0	42083
	[0.01,0.03)	0	31084	0	0	0	0	0	0	0	31084
	[0.03,0.05)	0	24108	0	0	0	0	0	0	0	24108
	[0.05,0.10)	0	55407	0	0	0	0	0	0	0	55407
	[0.10,0.50]	0	397265	0	0	0	0	0	0	0	397265
	Total	0	549947	0	0	0	0	0	0	0	549947
SINDI	[0]	0	0	0	0	0	0	0	0	0	0
	(0,0.01)	0	17615	0	0	0	0	0	0	0	17615
	[0.01,0.03)	0	20503	0	0	0	0	0	0	0	20503
	[0.03,0.05)	0	20826	0	0	0	0	0	0	0	20826
	[0.05,0.10)	0	57568	0	0	0	0	0	0	0	57568
	[0.10,0.50]	0	435766	0	0	0	0	0	0	0	435766
	Total	0	552278	0	0	0	0	0	0	0	552278

To facilitate downstream operations on genotype data, such as merging and meta-analysis, each dataset gets harmonized with modern reference data. The harmonization process is performed in two steps. First, using Genotype Harmonizer [2], the variants are strand-aligned with the 1000 Genomes Phase 3 Version 5 [4] variants. While some variants (A/C or G/T variants) may be removed due to strand ambiguity, if enough information exists, Genotype Harmonizer uses linkage disequilibrium (LD) patterns with nearby variants to accurately determine strand. This step will remove variants that it is unable to reconcile and maintains variants that are unique to the input data. The second step manually reconciles non-1000 Genomes variants with the human reference assembly GRCh37 [7]. This step will flag variants for removal that do not match an allele to the reference and variants that have only a single allele in the data file (0 for the other). Note that some monomorphic variants may be maintained in this process if there are two alleles in the data file and one of them matches a reference allele.

After harmonization, the data is loaded into a Hail [9] matrix table for downstream use. See Figure 2 for final variant counts by genotyping array.

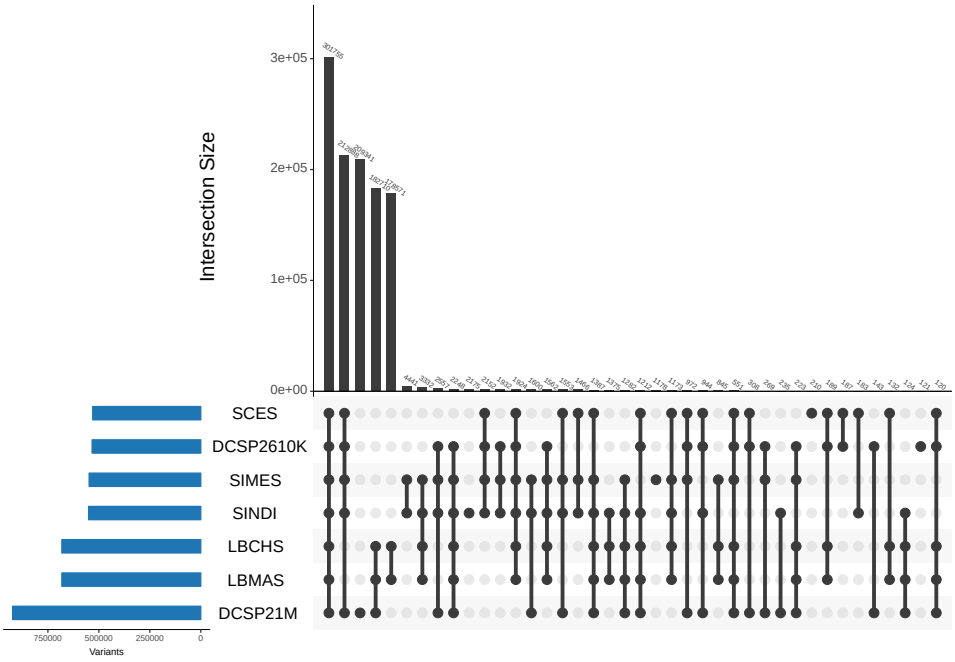


Figure 2: Variants remaining for analysis

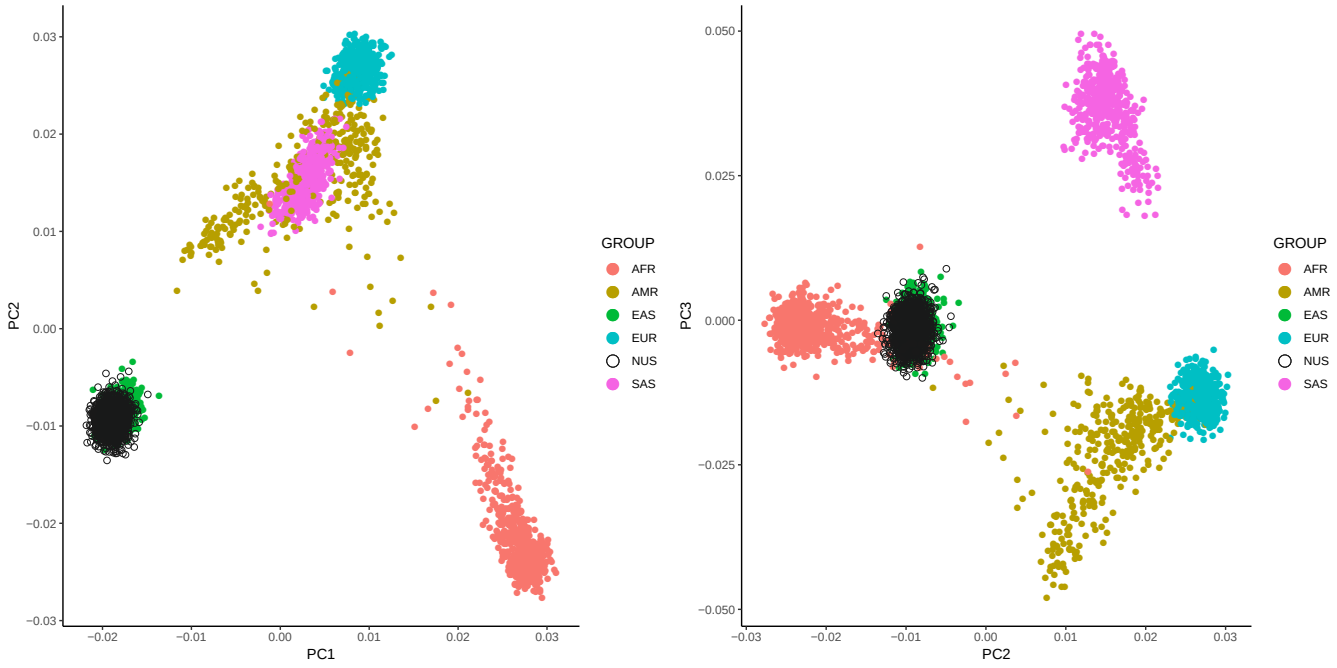
3 Sample QC

3.1 Ancestry Inference

Prior to association testing, it is useful to infer ancestry in relation to a modern reference panel representing the major human populations. While our particular sample QC process does not directly depend on this information, it is useful to downstream analysis when stratifying the calculation of certain variant statistics that are sensitive to population substructure (eg. Hardy Weinberg equilibrium). Additionally, ancestry inference may identify samples that do not seem to fit into a well-defined major population group, which would allow them to be flagged for removal from association testing.

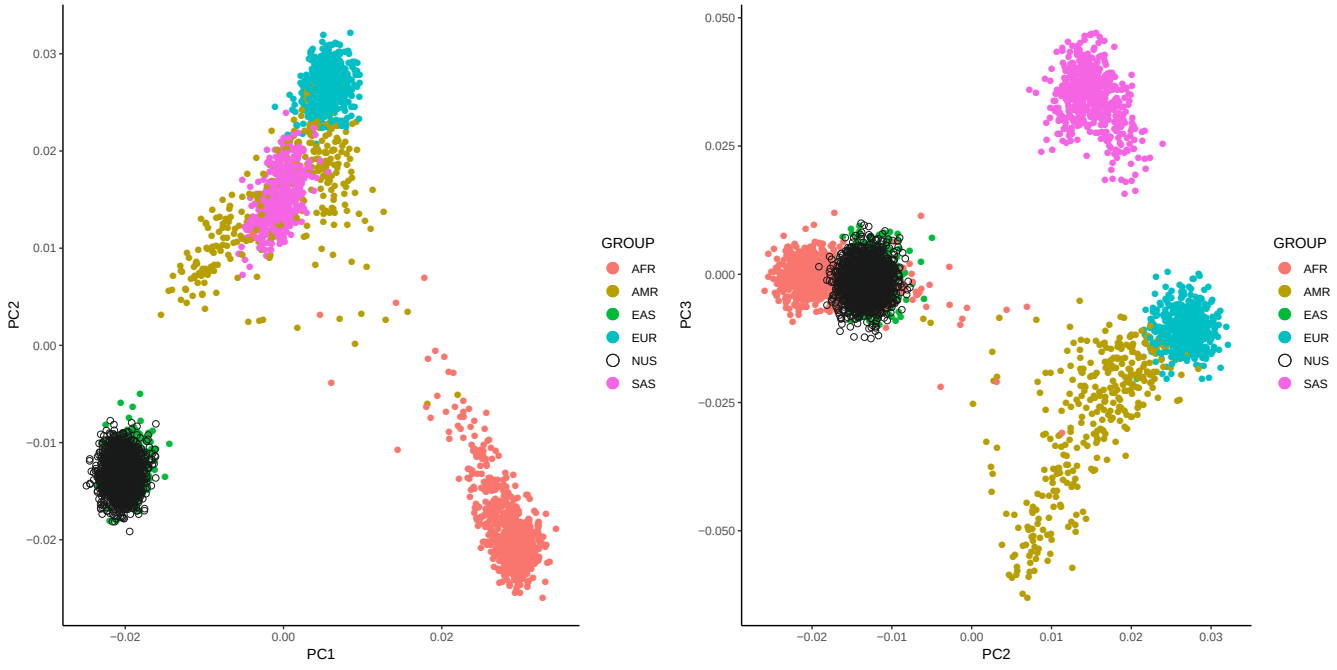
Initially, each array was merged with reference data. In this case, the reference used was the entire set of 2,504 1000 Genomes Phase 3 Version 5 [4] samples and our method restricted this merging to a set of 5,166 known ancestry informative SNPs. The merged data consisted of 5,432 DCSP21M, 2,745 DCSP2610K, 4,522 LBCHS, 4,521 LBMAS, 2,737 SCES, 2,736 SIMES and 2,743 SINDI variants. After merging, principal components (PCs) were computed using the PC-AiR [3] method in the GENESIS R package. This particular algorithm allows for the calculation of PCs that reflect ancestry in the presence of known or cryptic relatedness. The 1000 Genomes samples were forced into the 'unrelated' set and the PC-AiR algorithm was used to find the 'unrelated' samples from the array data. Then PCs were calculated on them and projected onto the remaining samples.

Figures 3, 4, 5, 6, 7, 8, and 9 display plots of the top three principal components along with the 1000 Genomes major population groups.



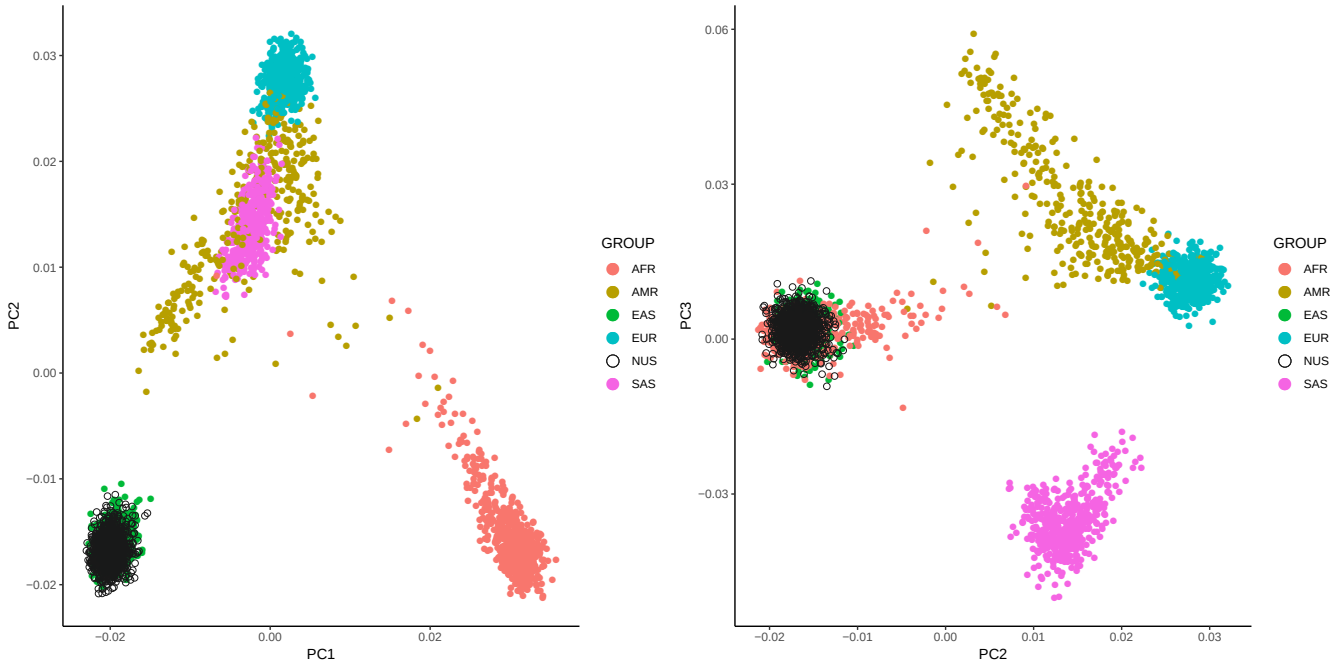
(a) PC1 vs. PC2 (b) PC2 vs. PC3

Figure 3: Principal components of ancestry for DCSP21M



(a) PC1 vs. PC2 (b) PC2 vs. PC3

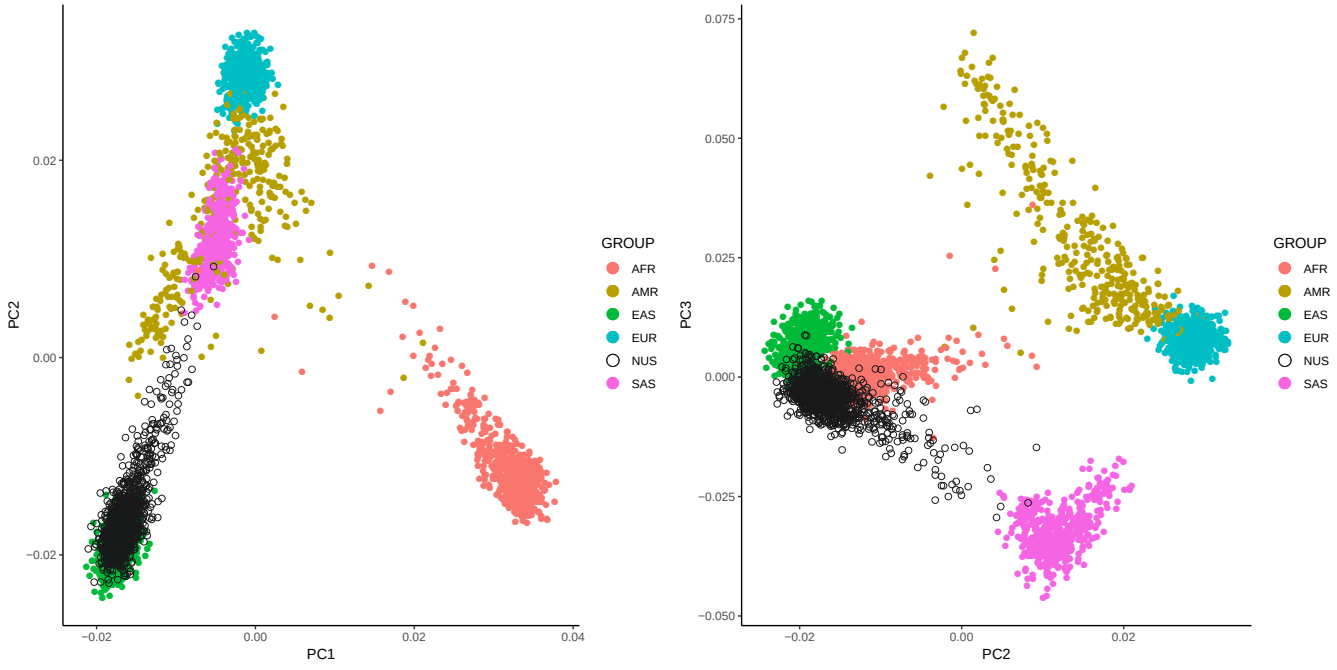
Figure 4: Principal components of ancestry for DCSP2610K



(a) PC1 vs. PC2

(b) PC2 vs. PC3

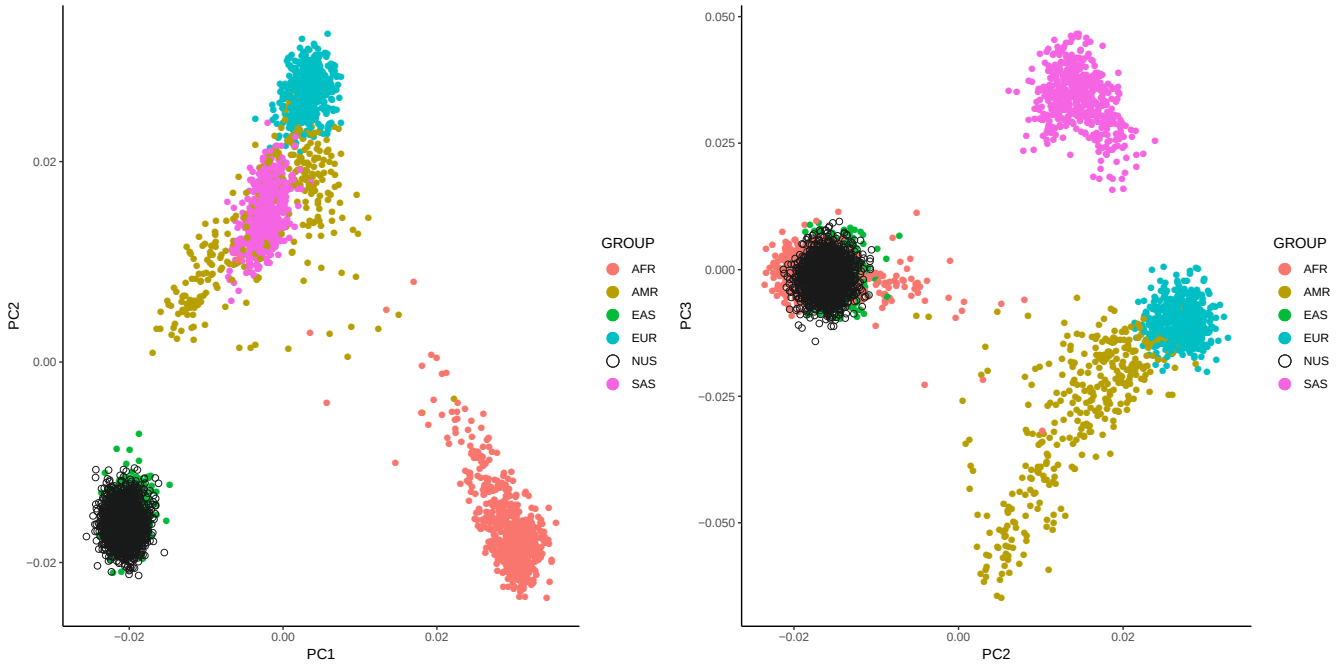
Figure 5: Principal components of ancestry for LBCHS



(a) PC1 vs. PC2

(b) PC2 vs. PC3

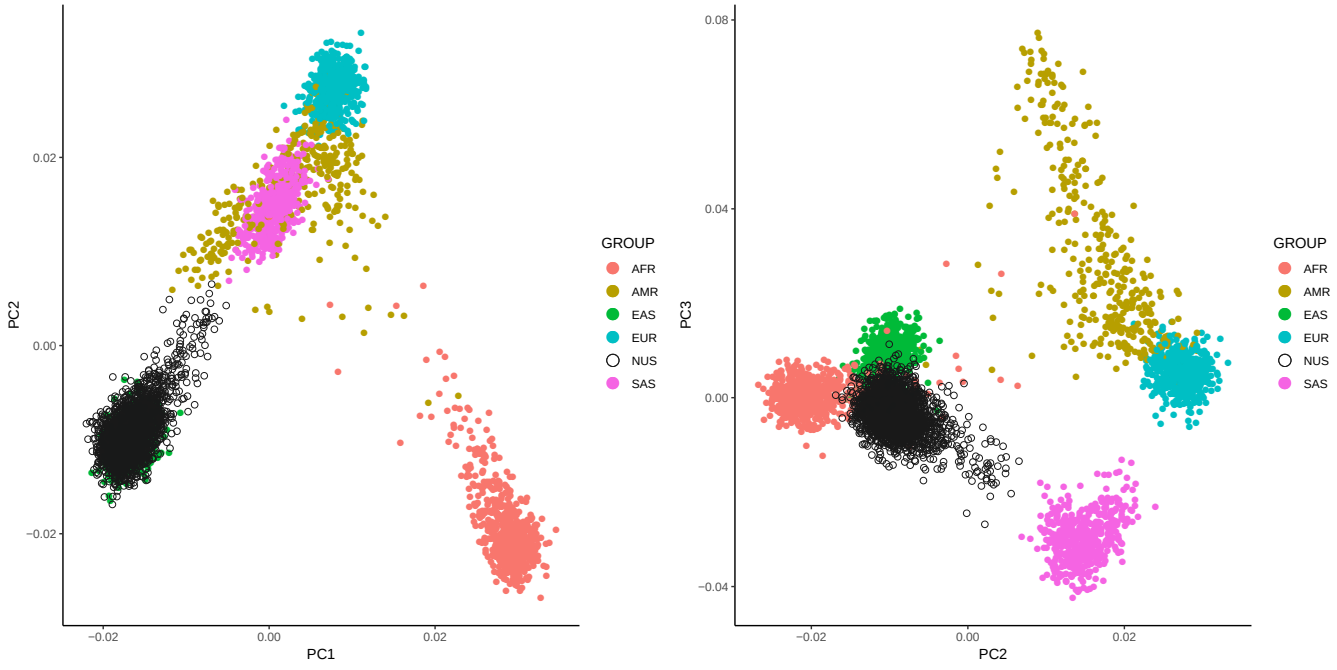
Figure 6: Principal components of ancestry for LBMAS



(a) PC1 vs. PC2

(b) PC2 vs. PC3

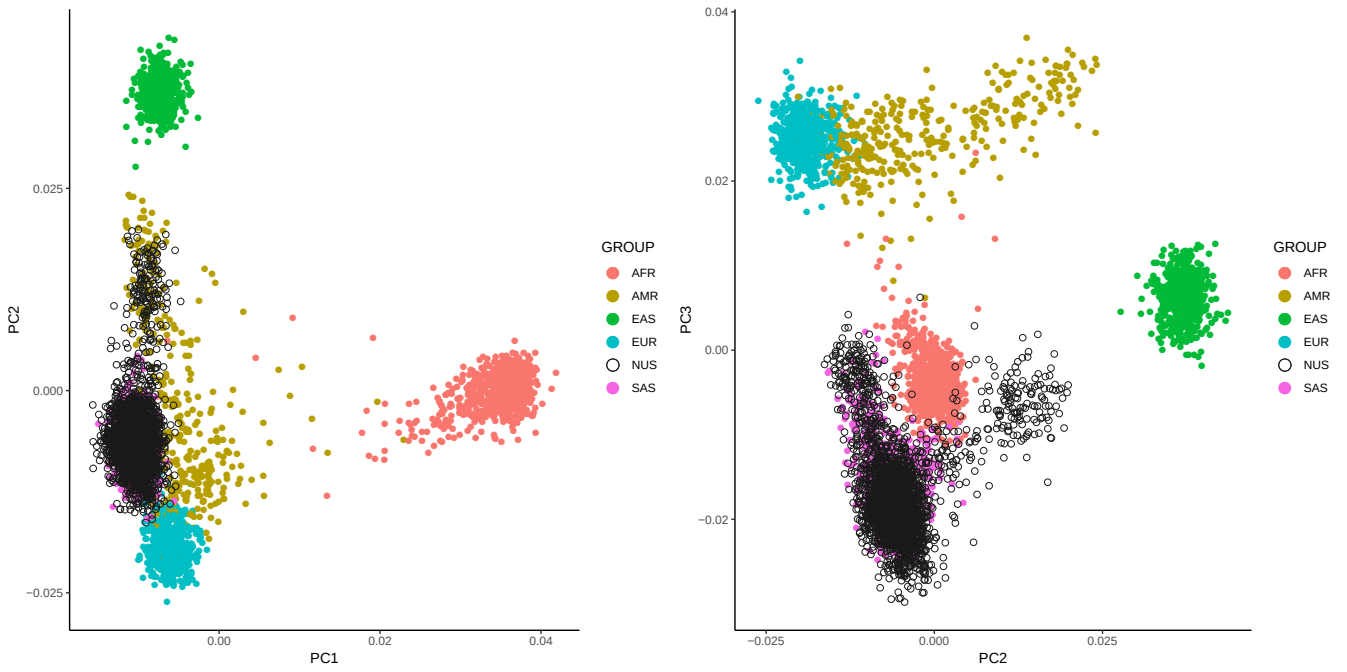
Figure 7: Principal components of ancestry for SCES



(a) PC1 vs. PC2

(b) PC2 vs. PC3

Figure 8: Principal components of ancestry for SIMES

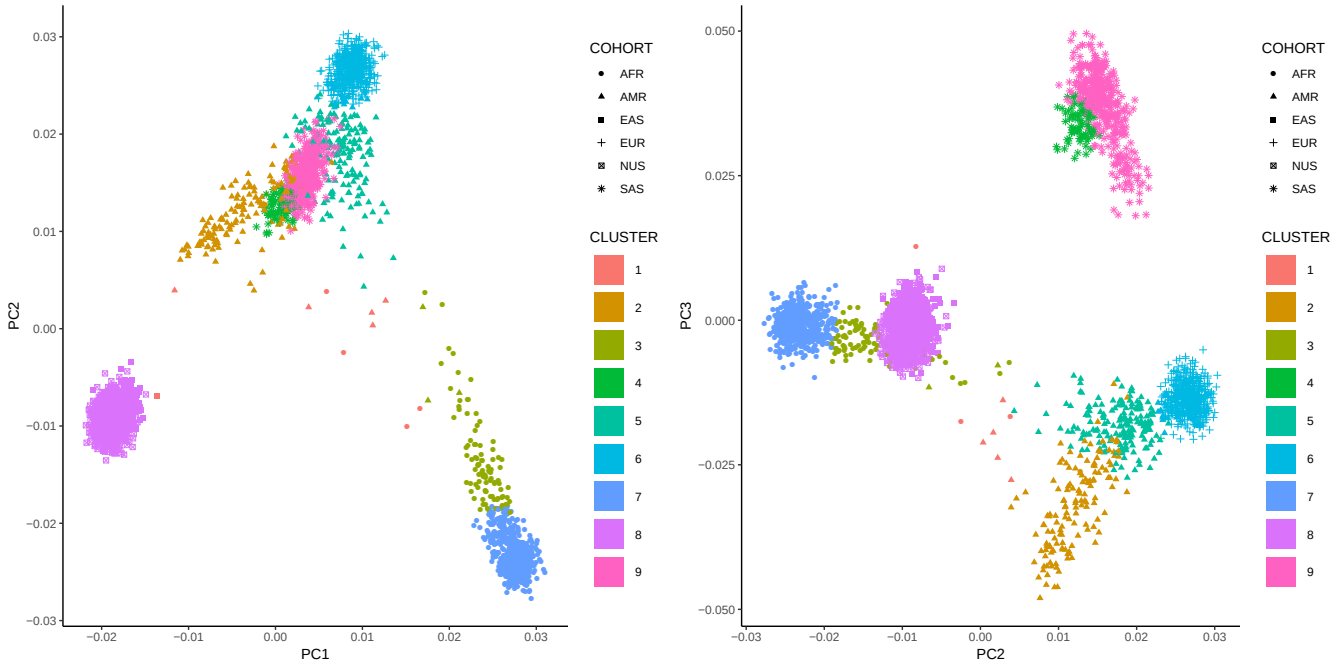


(a) PC1 vs. PC2

(b) PC2 vs. PC3

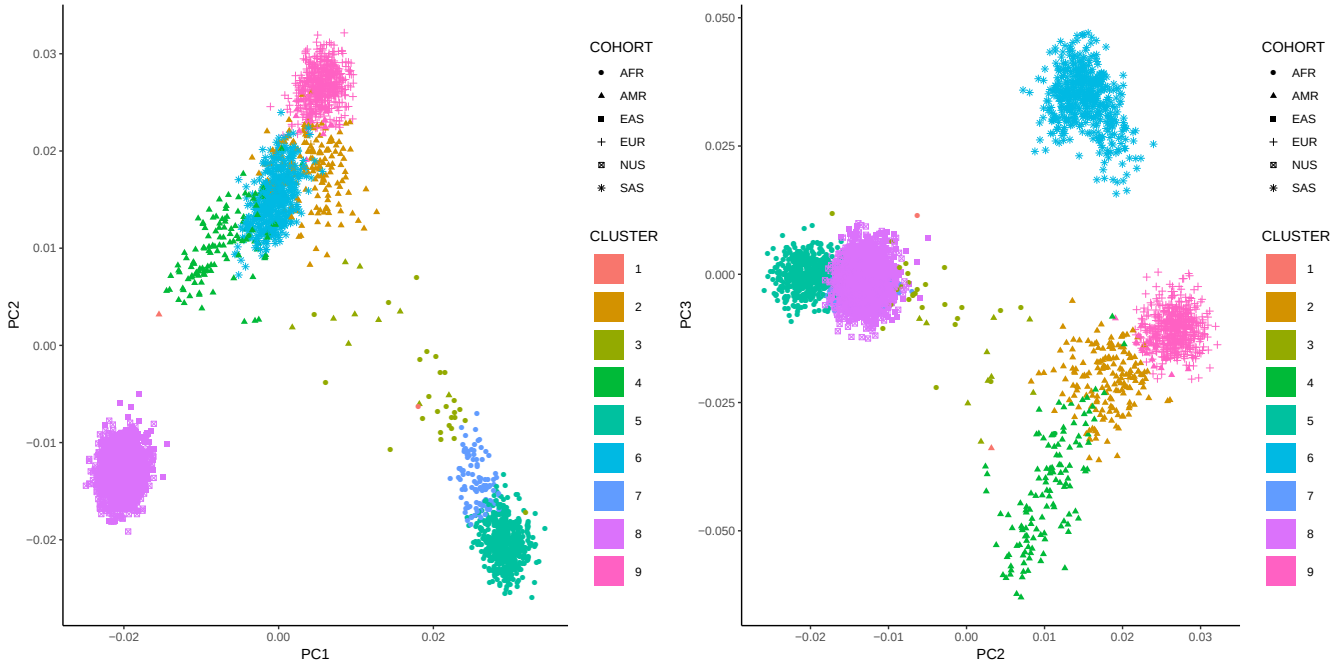
Figure 9: Principal components of ancestry for SINDI

Using the principal components of ancestry as features, we employed the signal processing software Klustakwik [5] to model the array as a mixture of Gaussians, identifying clusters, or population groups/subgroups. In order to generate clusters of sufficient size for statistical association tests, we used the first 3 principal components as features in the clustering algorithm. This number of PC's distinctly separates the five major 1000 Genomes population groups: AFR, AMR, EUR, EAS, and SAS. Figures 10, 11, 12, 13, 14, 15, and 16 clearly indicate the population structure in the datasets. In Klustakwik output, cluster 1 is always reserved for outliers, or samples that did not fit into any of the clusters found by the program. Upon further inspection, no samples were manually reinstated during this step.



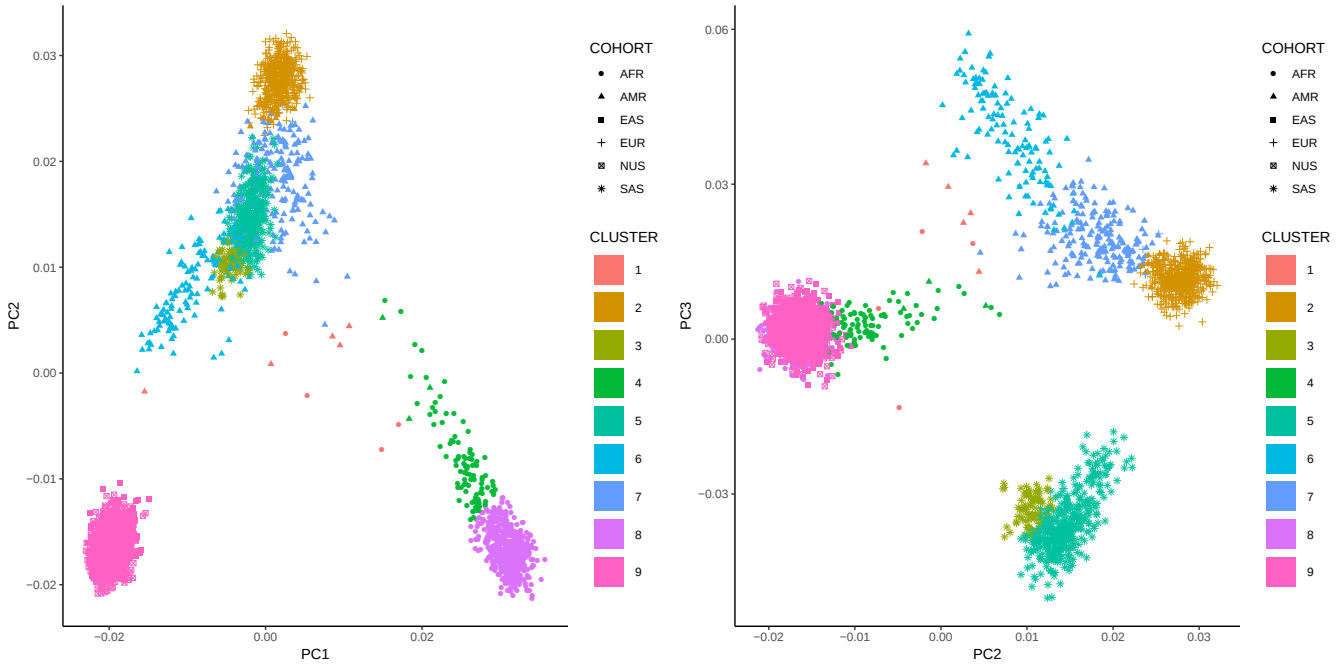
(a) PC1 vs. PC2 (b) PC2 vs. PC3

Figure 10: Population clusters for DCSP21M



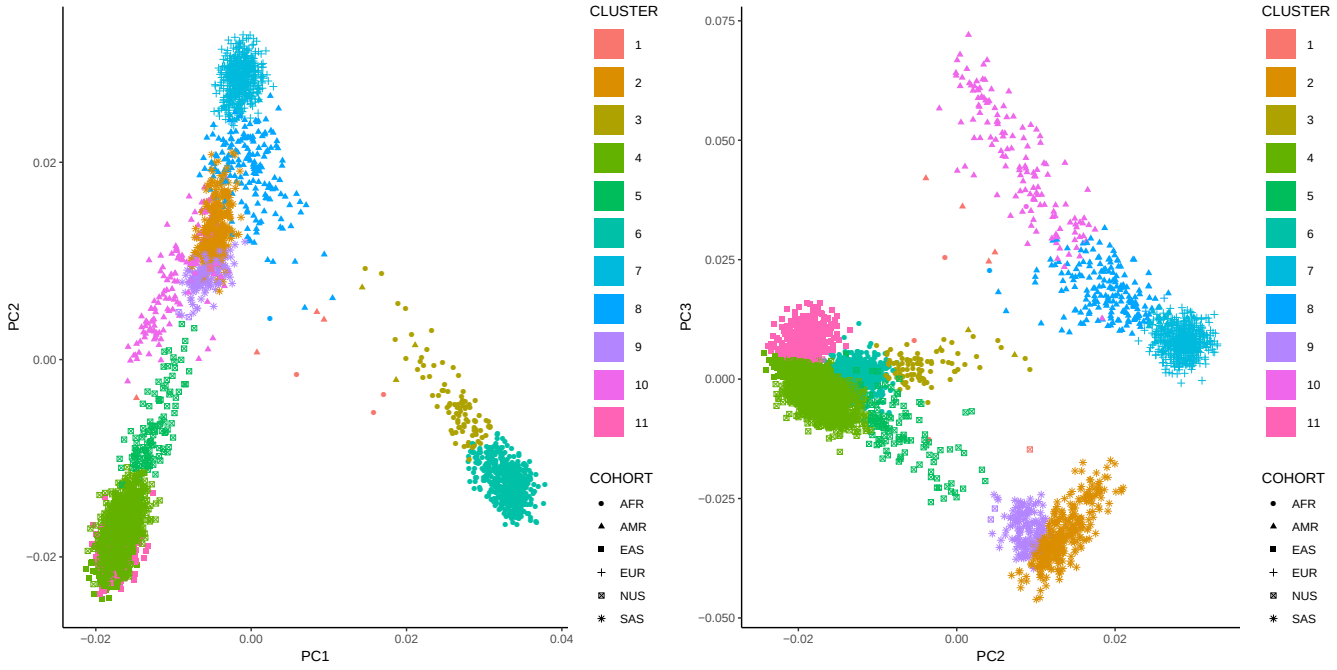
(a) PC1 vs. PC2 (b) PC2 vs. PC3

Figure 11: Population clusters for DCSP2610K



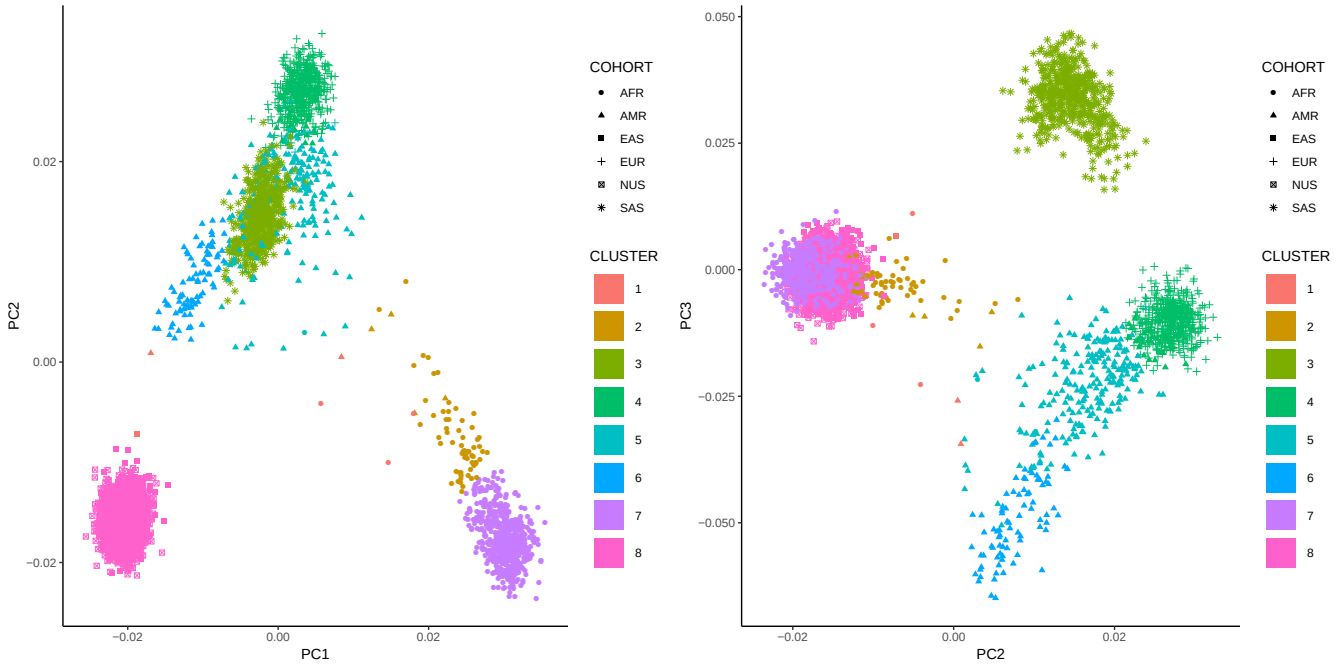
(a) PC1 vs. PC2 (b) PC2 vs. PC3

Figure 12: Population clusters for LBCHS



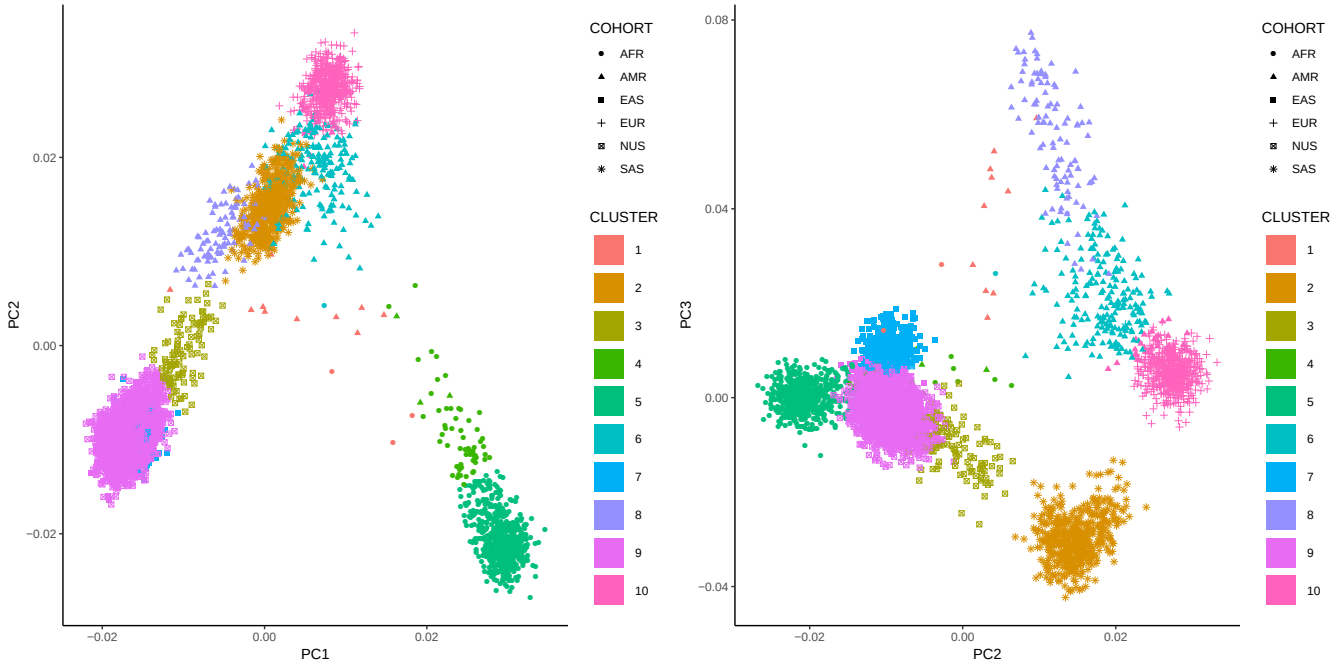
(a) PC1 vs. PC2 (b) PC2 vs. PC3

Figure 13: Population clusters for LBMAS



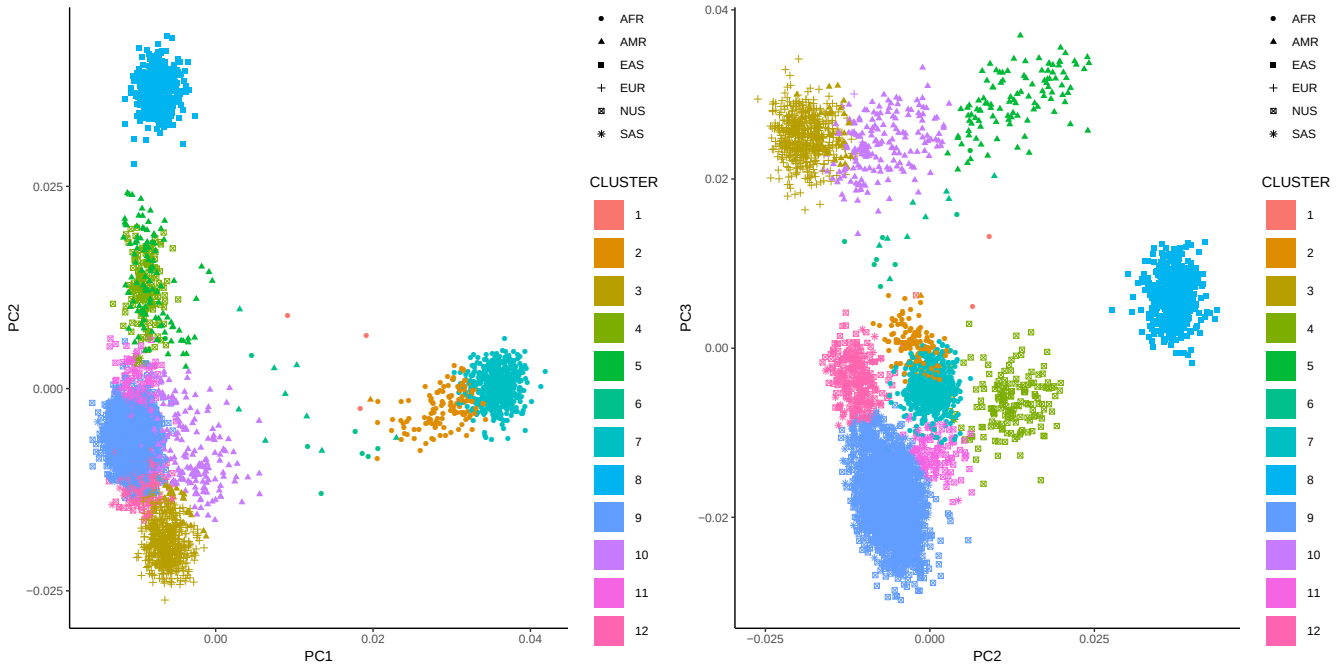
(a) PC1 vs. PC2 (b) PC2 vs. PC3

Figure 14: Population clusters for SCES



(a) PC1 vs. PC2 (b) PC2 vs. PC3

Figure 15: Population clusters for SIMES



(a) PC1 vs. PC2

(b) PC2 vs. PC3

Figure 16: Population clusters for SINDI

The resulting clusters are then combined with the nearest 1000 Genomes cohort. Table 2 describes the classification using this method. A final population assignment is determined by setting a hierarchy on the genotyping technologies (DCSP21M > DCSP2610K > LBCHS > LBMAS > SCES > SIMES > SINDI) and assigning each sample to the population determined using the highest technology.

Table 2: Inferred ancestry by dataset and cluster

	Population	Clusters	Samples
DCSP21M	EAS	8	1864
	Outliers	1	0
DCSP2610K	EAS	8	2087
	Outliers	1	0
LBCHS	EAS	9	1263
	Outliers	1	0
LBMAS	EAS	4,5,11	1185
	SAS	9	3
	Outliers	1	1
SCES	EAS	8	1889
	Outliers	1	0
SIMES	EAS	3,7,9	2542
	Outliers	1	0
SINDI	SAS	4,9,11,12	2537
	Outliers	1	1

Table 3: Final inferred ancestry

Population	Samples
EAS	10830
SAS	2540
Outliers	2

3.2 Duplicates and Excessive Sharing of Identity-by-Descent (IBD)

Sample pair kinship coefficients were determined using KING [8] relationship inference software, which offers a robust algorithm for relationship inference under population stratification. Prior to inferring relationships, we used Plink [1] to filter out non-autosomal, non-A/C/G/T, low callrate, and low minor allele frequency variants. Also, variants with positions in known high LD regions [6] and known Type 2 diabetes associated loci were

removed and an LD-pruned dataset was created. The specific filters that were used are listed below.

- --chr 1-22
- --snps-only just-acgt
- --exclude range ...
- --maf 0.01
- --geno 0.02
- --indep-pairwise 1000kb 1 0.2

After filtering there were 109,041 DCSP21M, 92,888 DCSP2610K, 101,404 LBCHS, 114,464 LBMAS, 92,909 SCES, 104,141 SIMES and 118,917 SINDI variants remaining.

In order to identify duplicate pairs of samples, a filter was set to $Kinship > 0.4$. There were no sample pairs identified as duplicate in the array data. Upon manual inspection, If the clinical data for any of the duplicate pairs was nearly identical (same date of birth, etc.), then the sample with the higher call rate was reinstated. If the clinical data did not match or a manual inspection was not performed, both samples were removed. In this case, no samples have been reinstated.

In addition to identifying duplicate samples, any single individual that exhibited kinship values indicating a 2nd degree relative or higher relationship with 10 or more others was flagged for removal. The relationship count indicated no samples that exhibited high levels of sharing identity by descent. Upon further inspection, no samples were manually reinstated during this step.

3.3 Sex Chromosome Check

Each array was checked for genotype / clinical data agreement for sex. There were no samples that were flagged as a 'PROBLEM' by Hail because it was unable to impute sex and there were no samples that were flagged for removal because the genotype based sex did not match their clinical sex. Upon further inspection, no samples were manually reinstated during this step.

3.4 Sample Outlier Detection

Each sample was evaluated for inclusion in association tests based on 10 sample-by-variant metrics (Table 4), calculated using Hail [9]. Note that for the metrics n_called and call_rate, only samples below the mean are filtered.

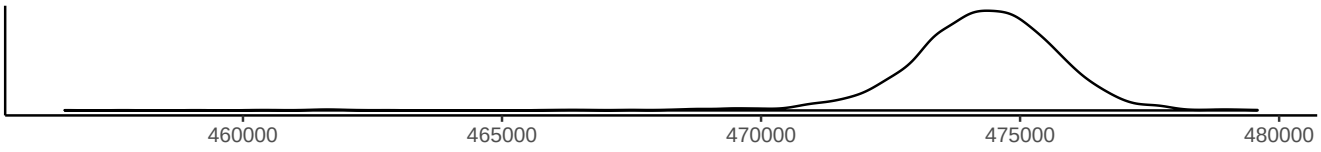
Table 4: Sample Metrics

n_non_ref	n_het + n_hom_var
n_het	Number of heterozygous variants
n_called	n_hom_ref + n_het + n_hom_var
call_rate	Fraction of variants with called genotypes
r_ti_tv	Transition/transversion ratio
het	Inbreeding coefficient
het_high	Inbreeding coefficient for variants with $MAF >= 0.03$
het_low	Inbreeding coefficient for variants with $MAF < 0.03$
n_hom_var	Number of homozygous alternate variants
r_het_hom_var	het/hom_var ratio across all variants

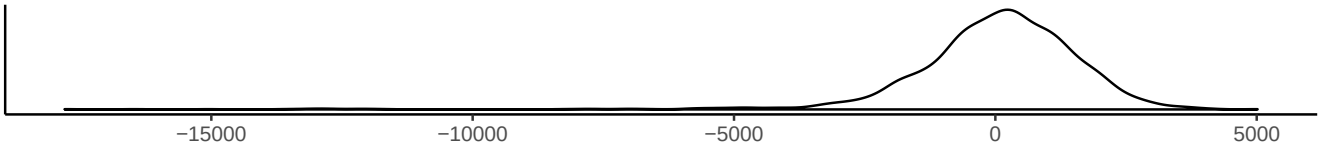
3.4.1 Principal Component Adjustment and Normalization of Sample Metrics

Due to possible population substructure, the sample metrics exhibit some multi-modality in their distributions. To evaluate more normally distributed data, we calculated principal component adjusted residuals of the metrics using the top 10 principal components (PCARM's). Figure 17 shows the n_non_ref metric for DCSP21M samples before and after adjustment.

Figure 17: Comparison of n_non_ref distributions before and after adjustment / normalization



(a) Original



(b) Adjusted

3.4.2 Individual Sample Metric Clustering

For outlier detection, we clustered the samples into Gaussian distributed subsets with respect to each PCARM using the software Klustakwik [5]. During this process, samples that did not fit into any Gaussian distributed set

of samples were identified and flagged for removal.

3.4.3 Principal Components of Variation in PCARM's

In addition to outliers along individual sample metrics, there may be samples that exhibit deviation from the norm across multiple metrics. In order to identify these samples, we calculated principal components explaining 95% of the variation in 8 of the 10 PCARMs combined. The adjusted residuals for metrics 'call_rate' and 'n_called' are characterized by long tails that lead to the maximum value, which is not consistent with the other metrics. In order to avoid excessive flagging of samples with lower, yet still completely acceptable, call rates, these metrics were left out of principal component calculation.

3.4.4 Combined PCARM Clustering

All samples were clustered into Gaussian distributed subsets along the principal components of the PCARM's, again using Klustakwik [5]. This effectively removed any samples that were far enough outside the distribution on more than one PCARM, but not necessarily flagged as an outlier on any of the individual metrics alone.

3.4.5 Plots of Sample Outliers

The distributions for each PCARM and any outliers (cluster = 1) found are shown in Figures 18, 19, 20, 21, 22, 23, and 24. Samples are labeled according to Table 5.

Table 5: Sample Legend for Outlier Plots

Grey	Clustered into Gaussian distributed subsets (not Flagged)
Orange	Flagged as outlier based on individual PCARM's
Blue	Flagged as outlier based on PC's of PCARM's
Green	Flagged as outlier for both methods

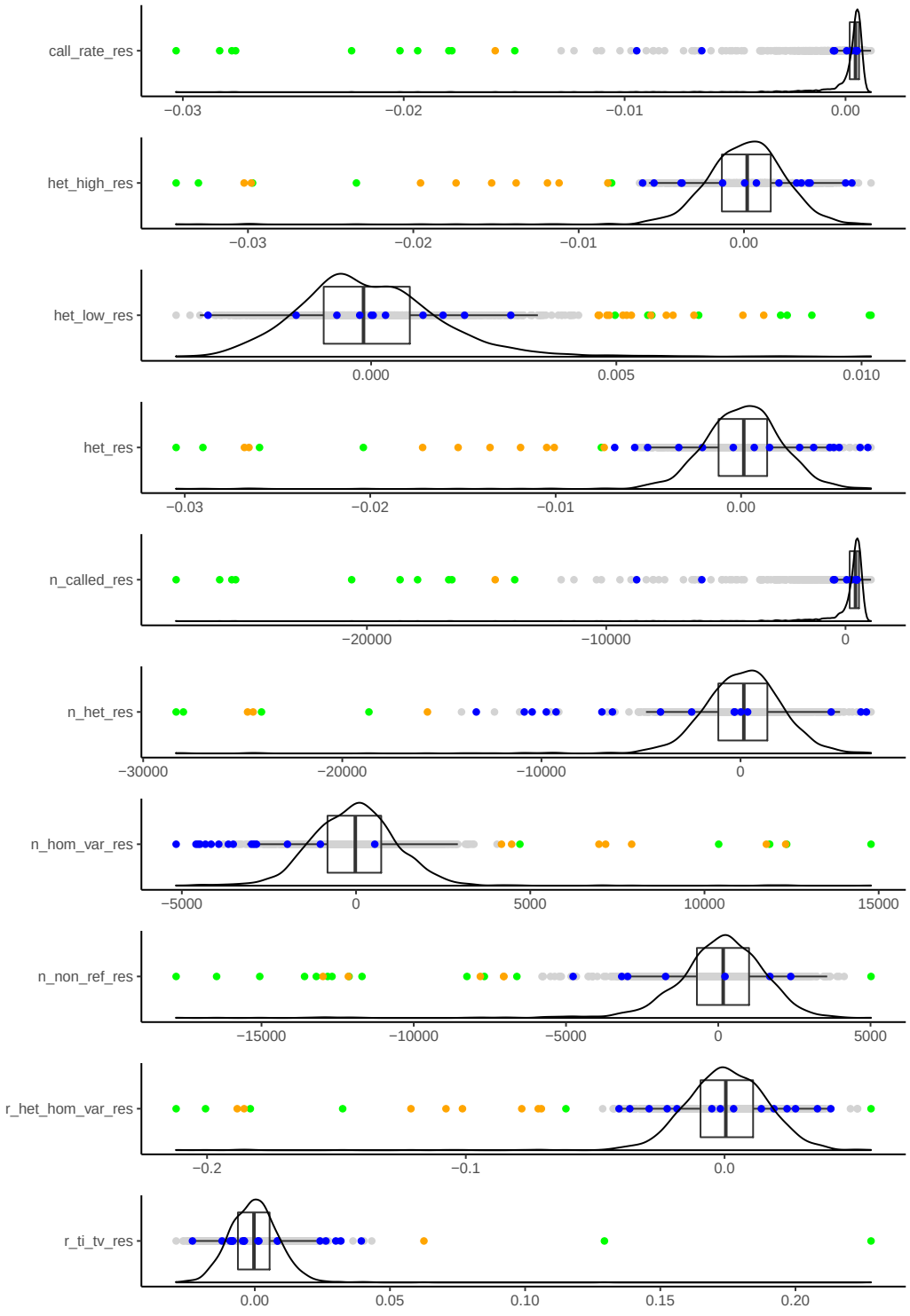


Figure 18: Adjusted sample metric distributions for DCSP21M

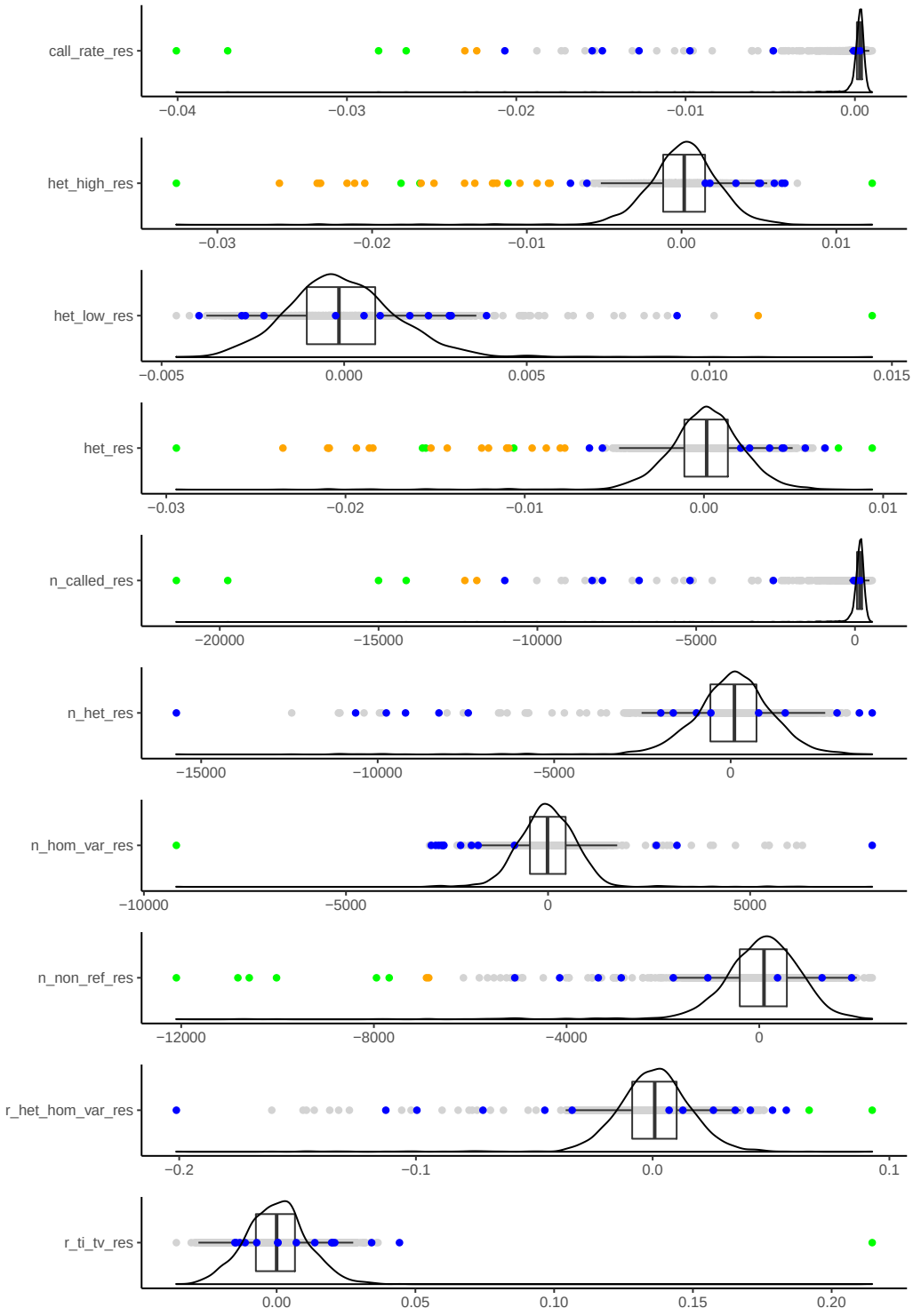


Figure 19: Adjusted sample metric distributions for DCSP2610K

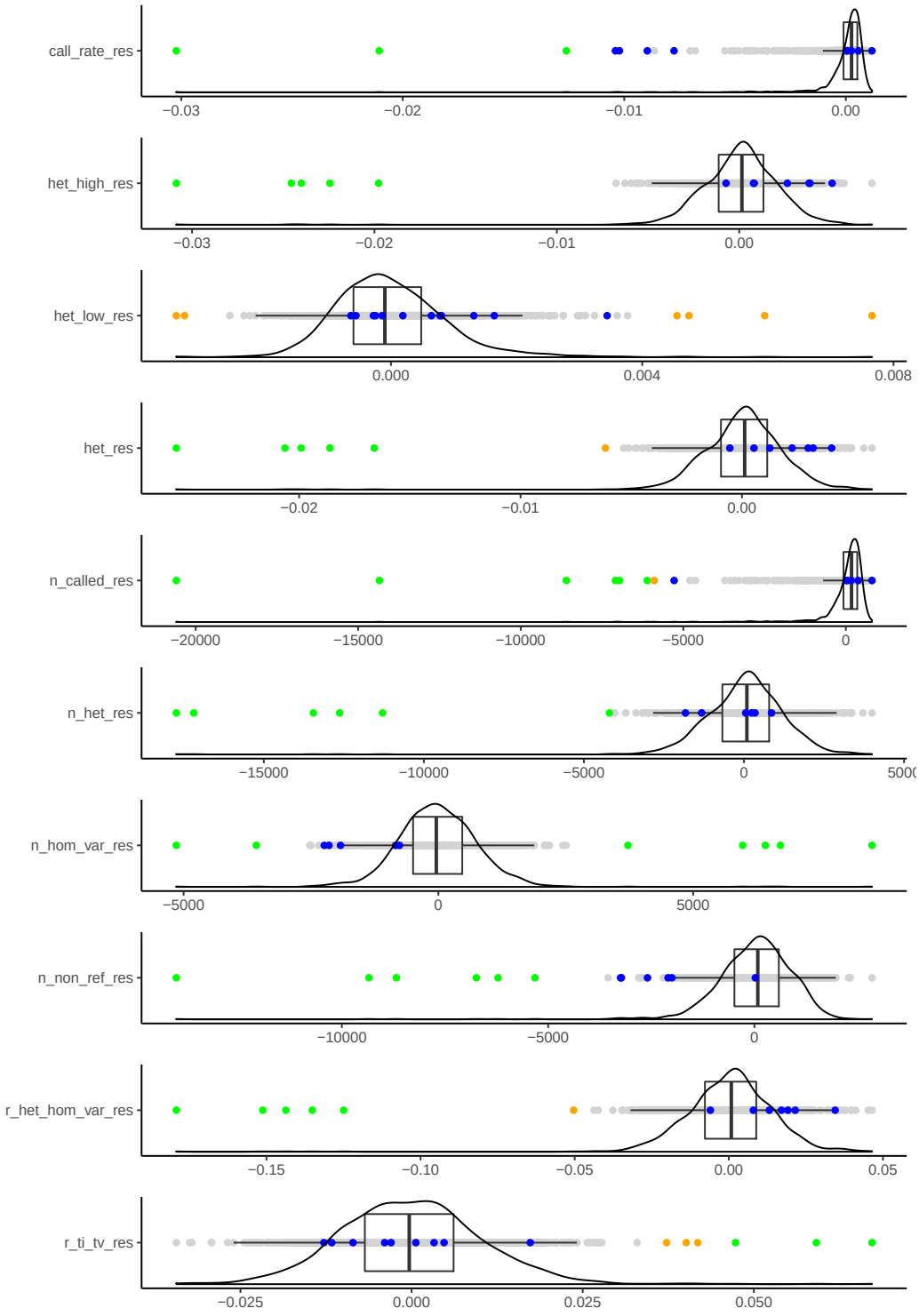


Figure 20: Adjusted sample metric distributions for LBCHS

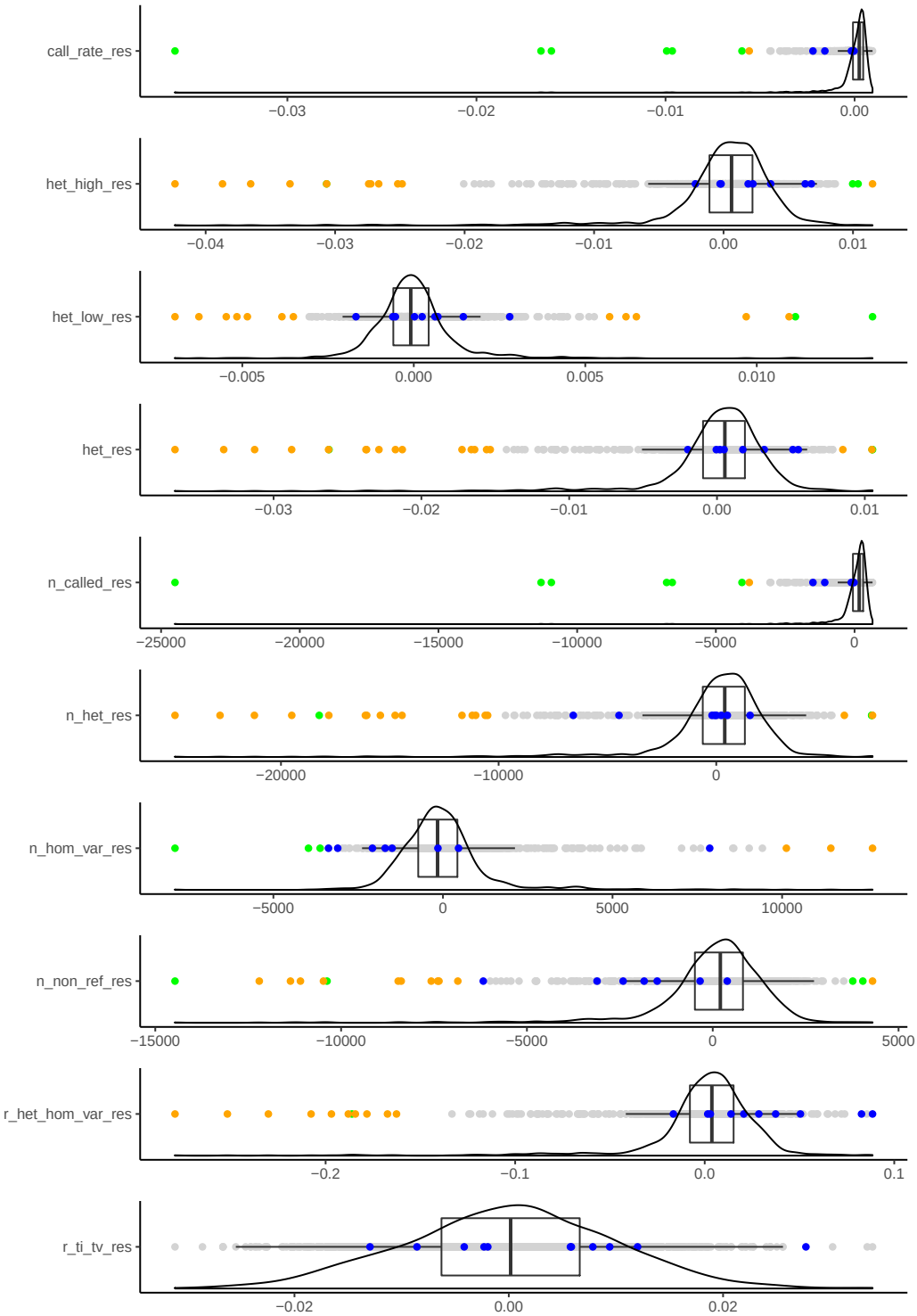


Figure 21: Adjusted sample metric distributions for LBMAS

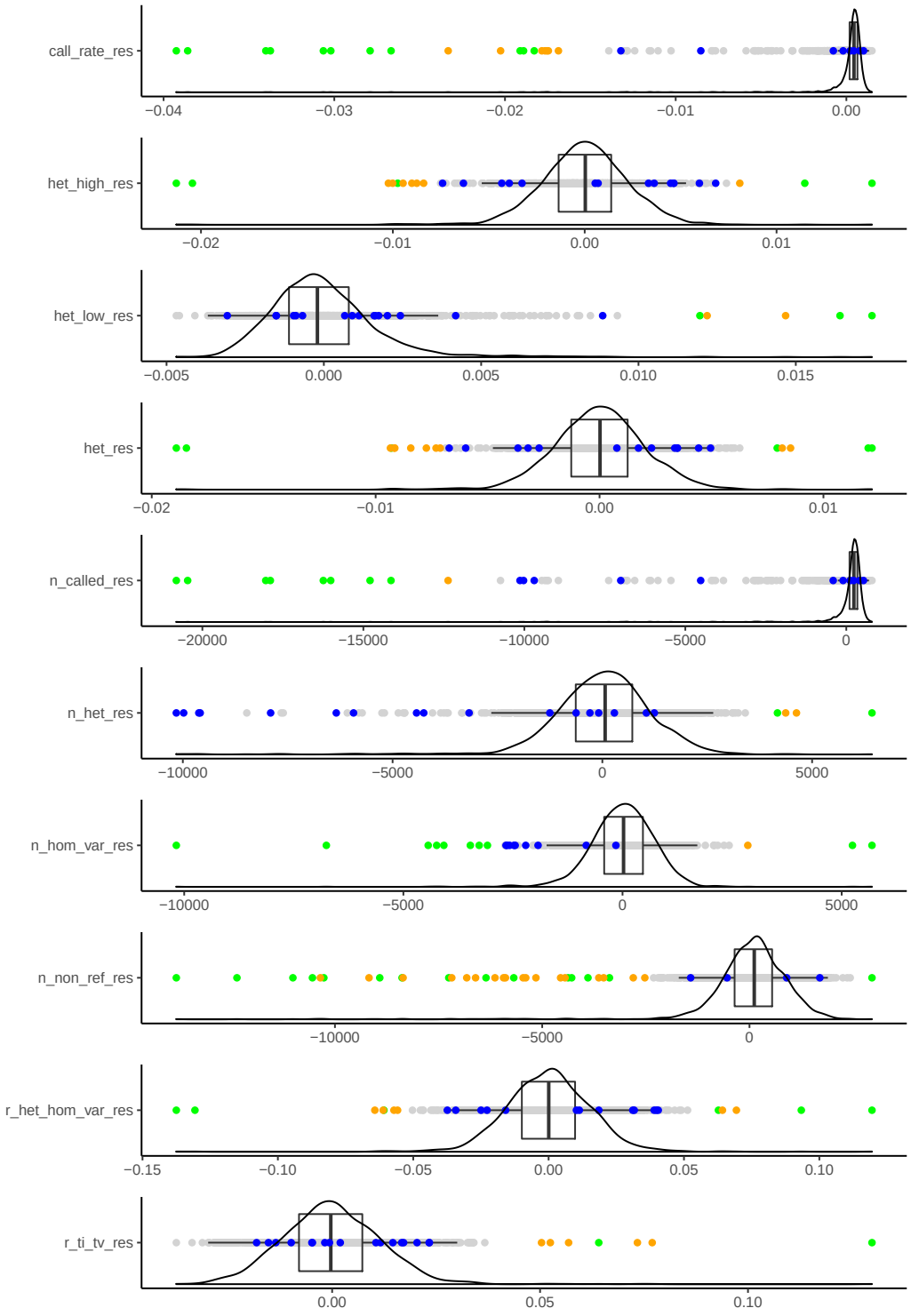


Figure 22: Adjusted sample metric distributions for SCES

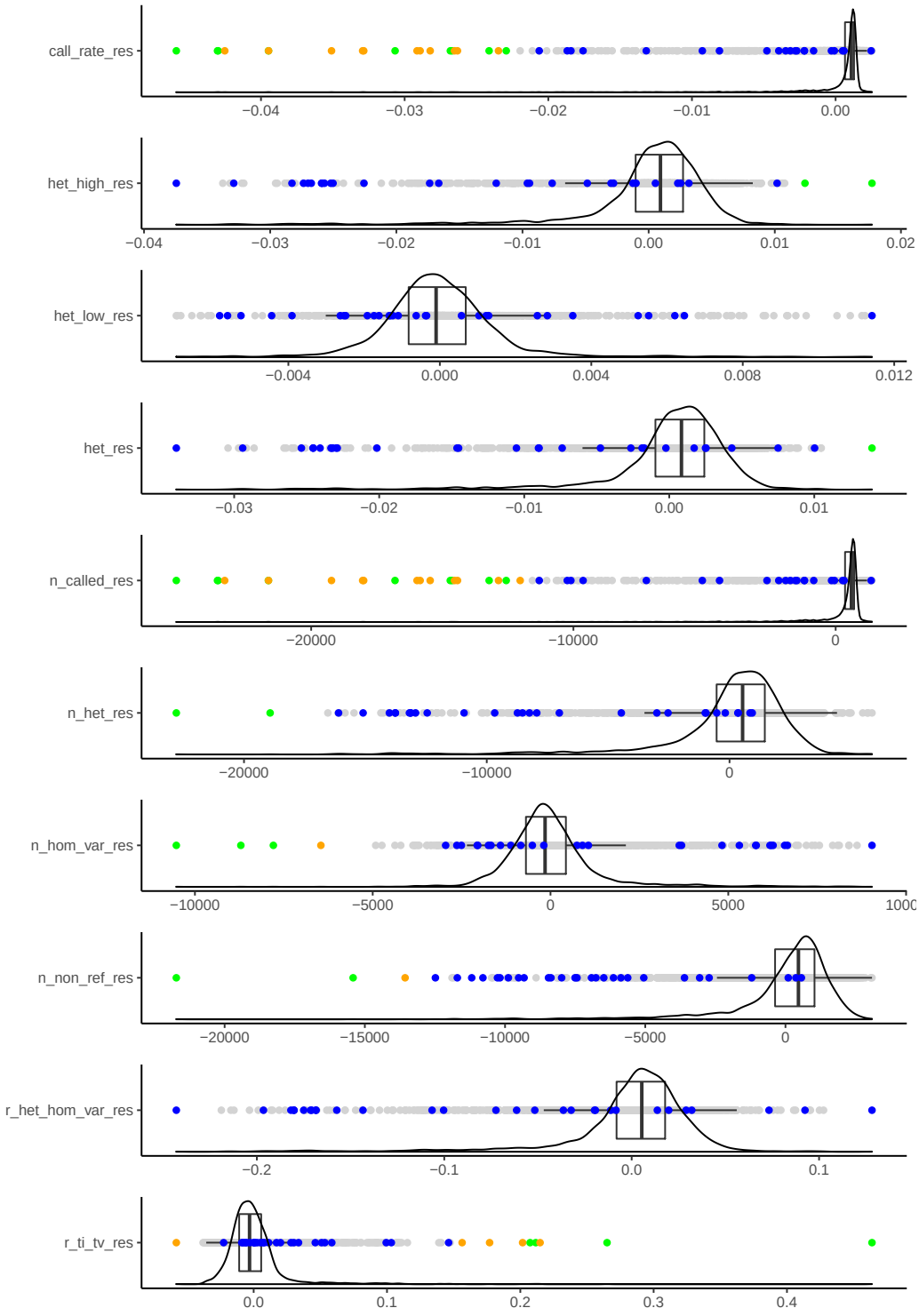


Figure 23: Adjusted sample metric distributions for SIMES

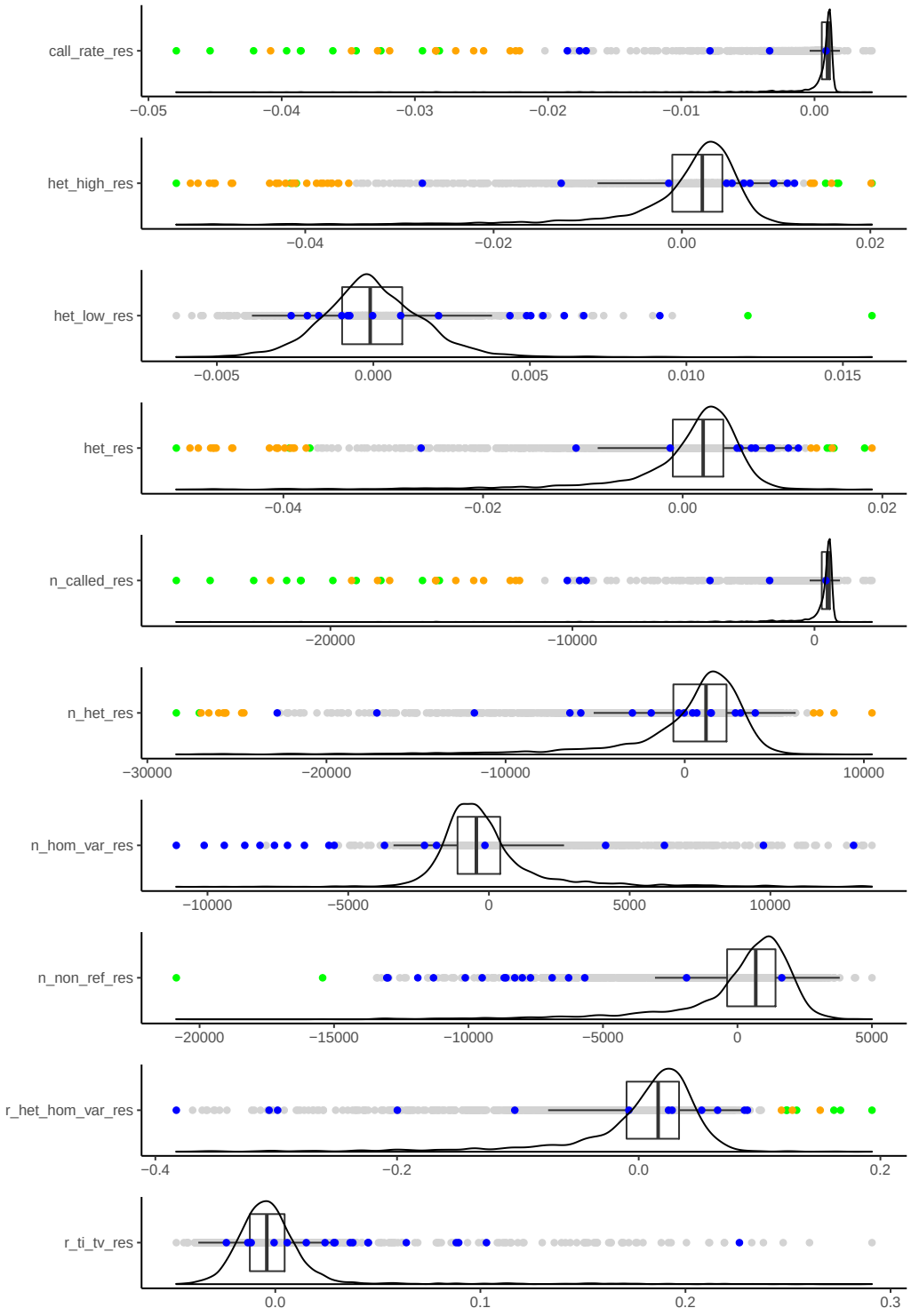


Figure 24: Adjusted sample metric distributions for SINDI

3.5 Summary of Sample Outlier Detection

Table 6 contains a summary of outliers detected by each method and across all genotyping technologies. Note that 'PCA(Metrics)' results from the clustering of the PCs of the 8 PCARM's combined, so 'Metrics + PCA(Metrics)' is the union of samples flagged by that method with samples flagged by each of the 10 individual metric clusterings. Figure 25 summarizes the samples remaining for analysis. Upon further inspection, no samples were manually reinstated during this step.

Table 6: Samples flagged for removal

	DCSP21M	DCSP2610K	LBCHS	LBMAS	SCES	SIMES	SINDI	Total
call_rate	21	17	12	12	26	41	30	159
het_high	29	33	12	22	26	30	49	201
het_low	34	16	18	23	21	30	18	160
het	29	33	13	28	28	30	39	200
n_called	21	17	13	12	20	42	30	155
n_het	23	15	12	28	21	30	30	159
n_hom_var	27	15	12	14	20	31	18	137
n_non_ref	25	17	12	22	37	31	18	162
r_het_hom_var	28	15	13	21	25	30	21	153
r_ti_tv	21	15	15	11	24	35	18	139
PCA(Metrics)	20	15	12	11	19	30	18	125
Metrics+PCA(Metrics)	44	36	22	40	42	47	60	291
Extreme Missingness	0	0	0	0	0	0	0	0
Duplicates	0	0	0	0	0	0	0	0
Cryptic Relatedness	0	0	0	0	0	0	0	0
Sexcheck	0	0	0	0	0	0	0	0
Ancestry Outlier	0	0	0	0	0	0	0	0
Total	46	38	24	42	44	49	62	293

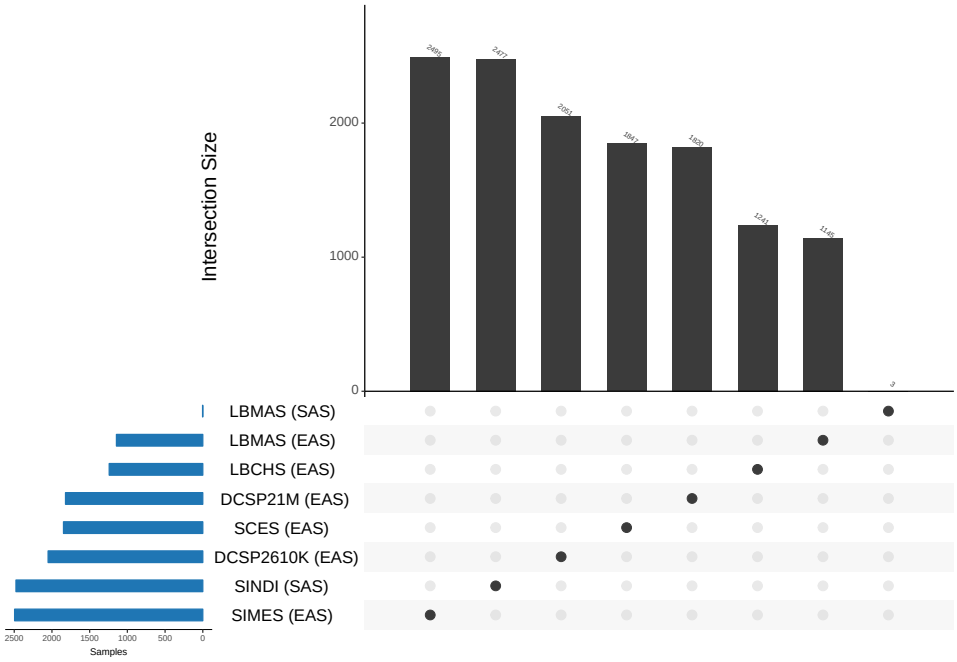


Figure 25: Samples remaining for analysis

4 Variant QC

Variant quality was assessed using call rate and Hardy Weinberg equilibrium (HWE). We calculate HWE using controls only within any of 4 major ancestral populations; EUR, AFR, SAS and EAS. There must have been at least 100 samples in a population to trigger a filter. This conservative approach minimizes the influence from admixture in other population groups. This procedure resulted in flagging 4,349 DCSP21M, 1,169 DCSP2610K, 3,511 LBCHS, 3,312 LBMAS, 2,220 SCES, 2,093 SIMES and 2,467 SINDI variants for removal. Figure 26 shows the number of variants remaining for analysis after applying filters.

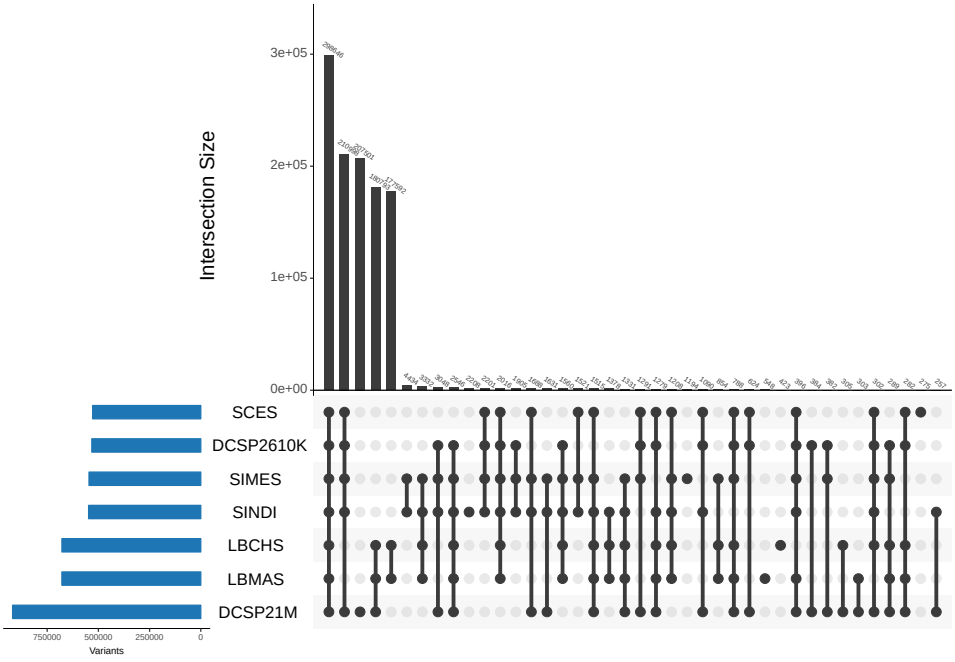


Figure 26: Variants remaining for analysis

5 Acknowledgements

We would like to acknowledge the following people for their significant contributions to this work.

Ryan Koesterer

Jason Flannick

Marcin von Grotthuss

6 References

[1] Plink1.9, https://www.cog-genomics.org/plink2.

[2] Deelan P, Bonder MJ, Joeri van der Velde K, Westra HJ, Winder E, Hendriksen D, Franke L, Swertz MA. Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Research Notes; 2014. 7:901. doi:10.1186/1756-0500-7-901. https://github.com/molgenis/systemsgenetics/wiki/Genotype-Harmonizer.

[3] Conomos MP. GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness, https://www.rdocumentation.org/packages/GENESIS/versions/2.2.2.

[4] 1000 Genomes Phase 3 v5, https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html.

[5] Klustakwik, http://klustakwik.sourceforge.net/.

[6] http://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD).

[7] https://www.ncbi.nlm.nih.gov/grc/human/data?asm=GRCh37.

[8] http://people.virginia.edu/~wc9c/KING/.

[9] Seed C, Bloemendal A, Bloom JM, Goldstein JI, King D, Poterba T, Neale BM. Hail: An Open-Source Framework for Scalable Genetic Data Analysis. In preparation. https://github.com/hail-is/hail.

[10] Gilbert C, Ruebenacker O, Koesterer R, Massung J, Flannick J. Loamstream. loamstream 1.4-SNAPSHOT (1.3-329-g0da8aac) branch: cg-h2-replacement commit: 0da8aac39f7f23cc5442b14a6ee77a767e1d9e35 built on: 2019-09-30T14:53:18.723Z. https://github.com/broadinstitute/dig-loam-stream.

[11] Koesterer R, Gilbert C, Ruebenacker O, Massung J, Flannick J. AMP-DCC Data Analysis Pipeline. dig-loam-2.5.26. https://github.com/broadinstitute/dig-loam.