AMP-DCC Quality Control Report

BioMe

05/16/2019 (02:08)

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 BioMe dataset. We received genotypes for 13,039 unique samples distributed across 3 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 3 genotype arrays, along with their intersection sizes.

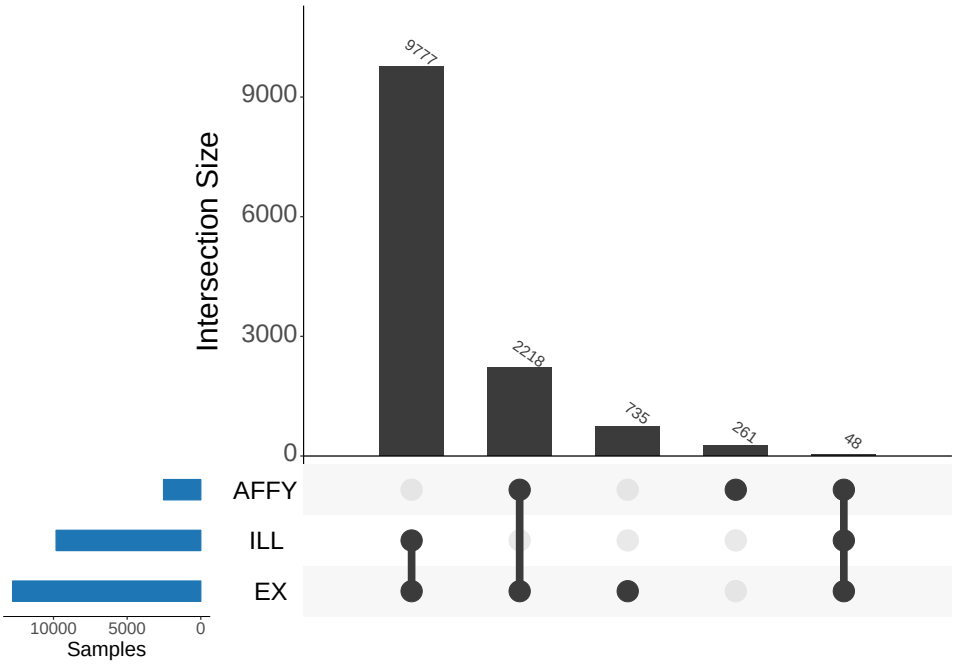


Figure 1: Samples distributed by genotyping array

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 = Mitochodrial 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
EX	[0]	0	33624	933	0	0	4	20	0	98	34561
	(0,0.01)	0	156059	3080	1	0	130	47	0	518	159270
	[0.01,0.03)	0	13255	261	0	0	35	25	0	66	13551
	[0.03,0.05)	0	4221	85	0	0	14	9	0	29	4320
	[0.05,0.10)	0	4942	101	0	0	12	13	0	21	5055
	[0.10,0.50]	0	21801	468	0	0	9	19	0	60	22278
	Total	0	233902	4928	1	0	204	133	0	792	239035
ILL	[0]	4	4781	267	0	2	0	3	0	29	5054
	(0,0.01)	96	160384	3293	0	29	127	47	0	1559	163929
	[0.01,0.03)	44	32966	770	0	15	38	24	0	911	33833
	[0.03,0.05)	20	32201	723	0	11	18	10	0	681	32973
	[0.05,0.10)	66	81626	1780	0	37	14	13	0	1438	83523
	[0.10,0.50]	364	534011	12909	0	255	13	19	4	12196	547552
	Total	594	845969	19742	0	349	210	116	4	16814	866864
AFFY	[0]	0	236	442	40	0	1	0	0	0	719
	(0,0.01)	0	7493	919	42	0	0	0	0	0	8454
	[0.01,0.03)	0	39413	2084	40	0	0	0	0	0	41537
	[0.03,0.05)	0	55249	1943	47	0	1	0	0	0	57240
	[0.05,0.10)	0	121326	4283	36	0	5	0	0	0	125650
	[0.10,0.50]	0	614104	26040	173	0	107	0	0	0	640424
	Total	0	837821	35711	378	0	114	0	0	0	874024

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.

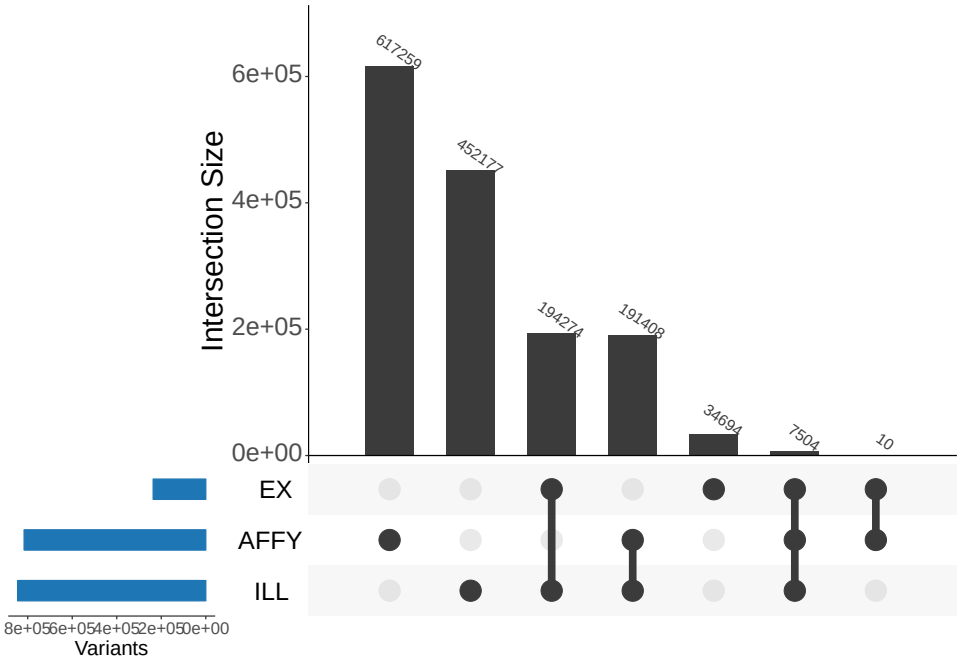


Figure 2: Variants remaining for analysis

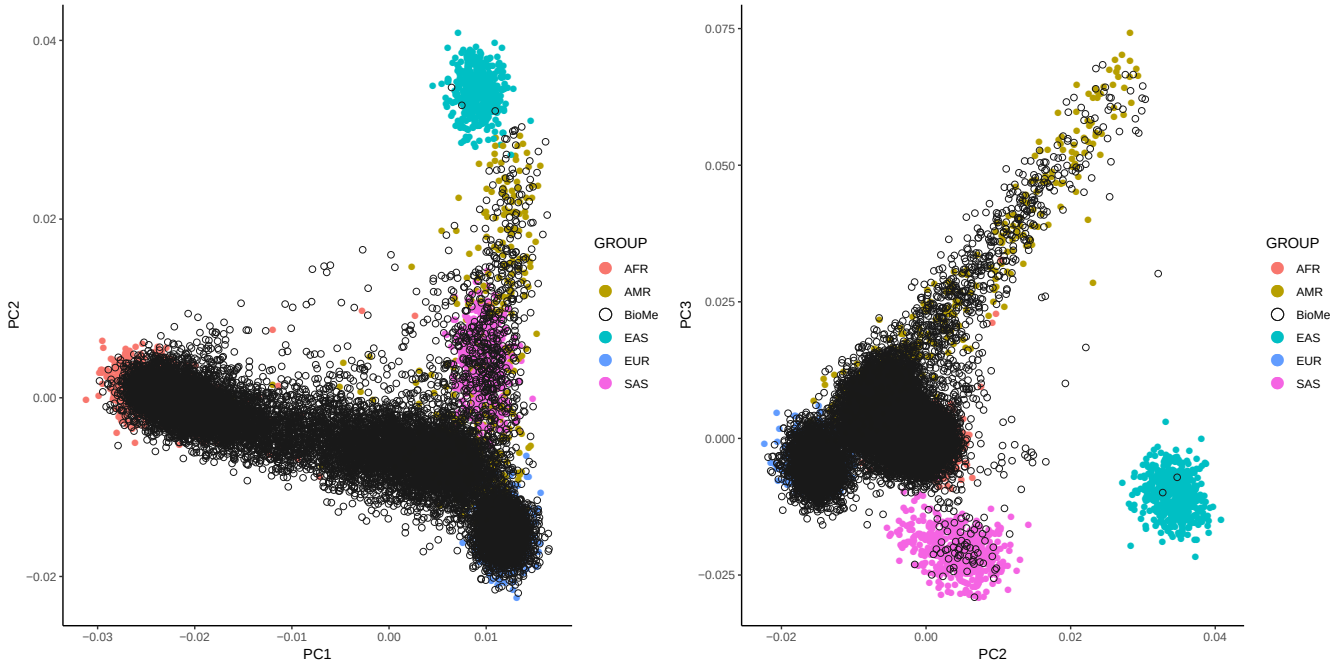
3 Sample QC

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 2,260 EX, 4,758 ILL and 1,668 AFFY 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.

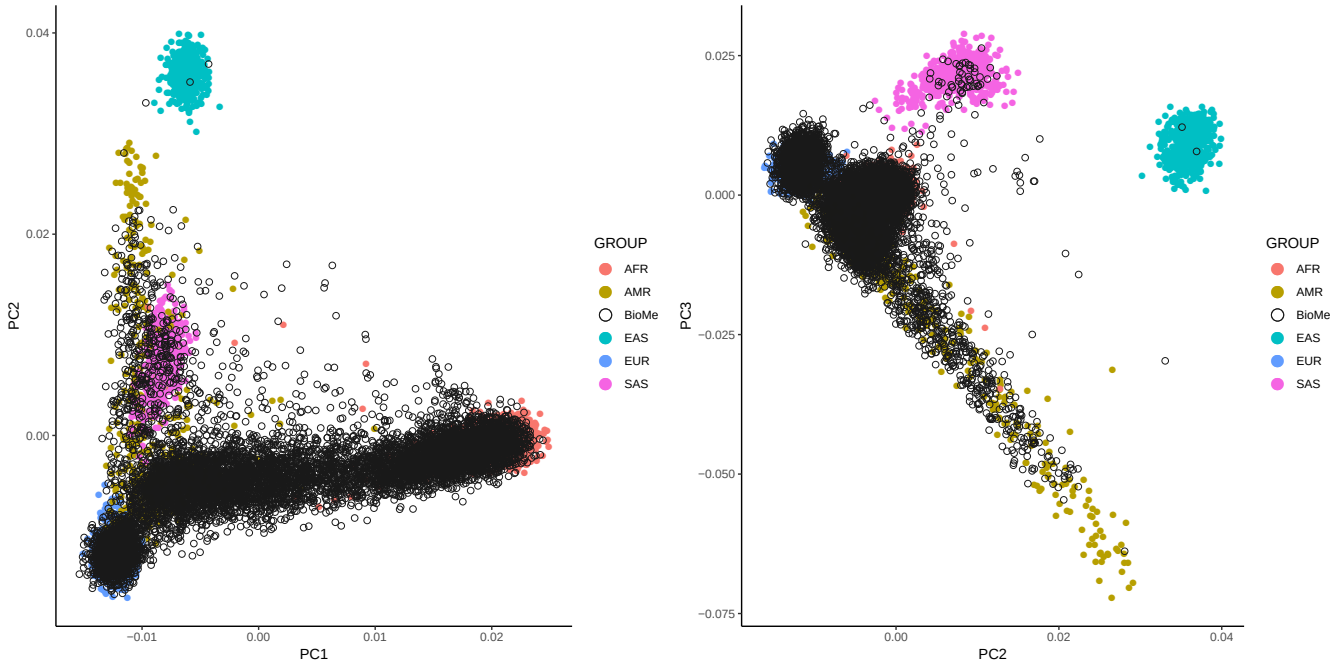
Figures 3, 4, and 5 display plots of the top three principal components along with the 1000 Genomes major population groups.



(a) PC1 vs. PC2

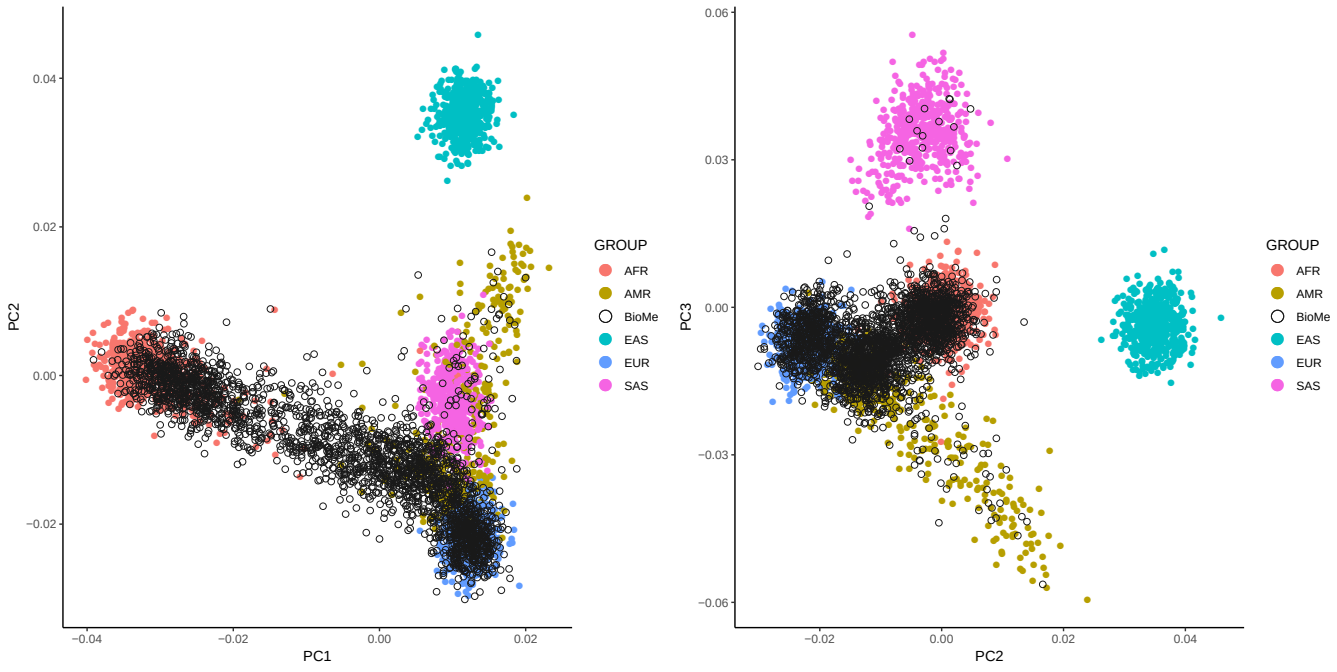
(b) PC2 vs. PC3

Figure 3: Principal components of ancestry for EX



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

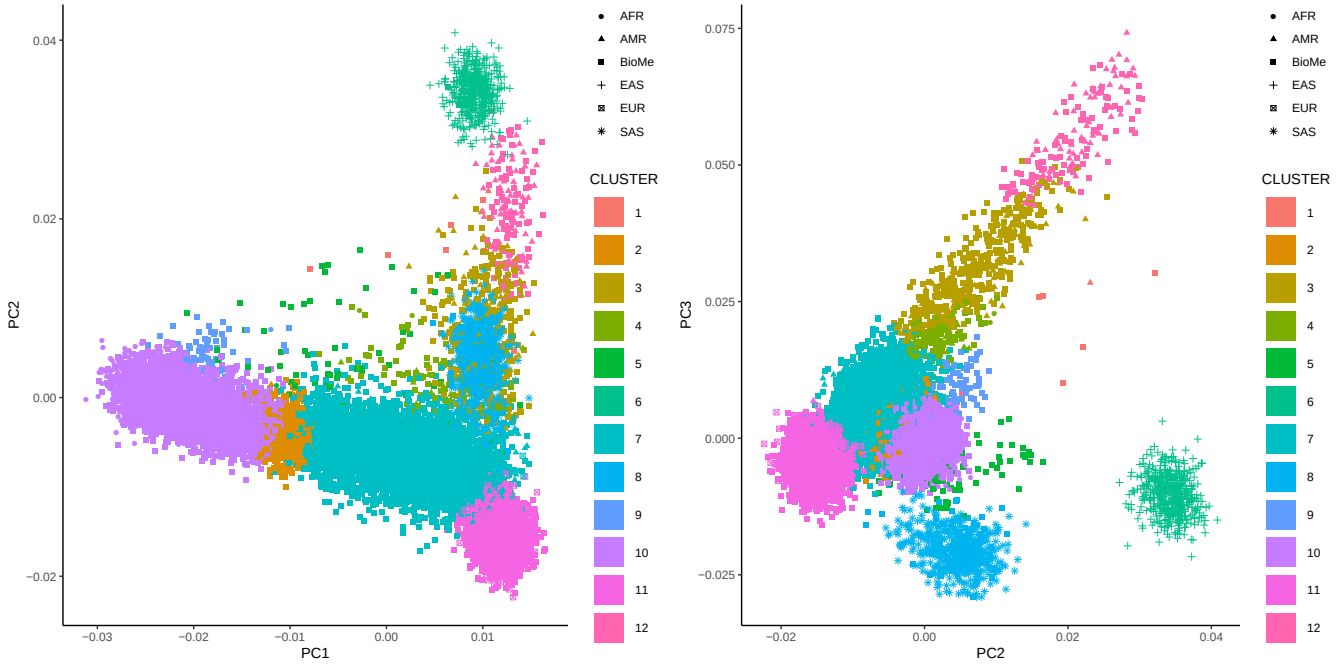
Figure 4: Principal components of ancestry for ILL



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

Figure 5: Principal components of ancestry for AFFY

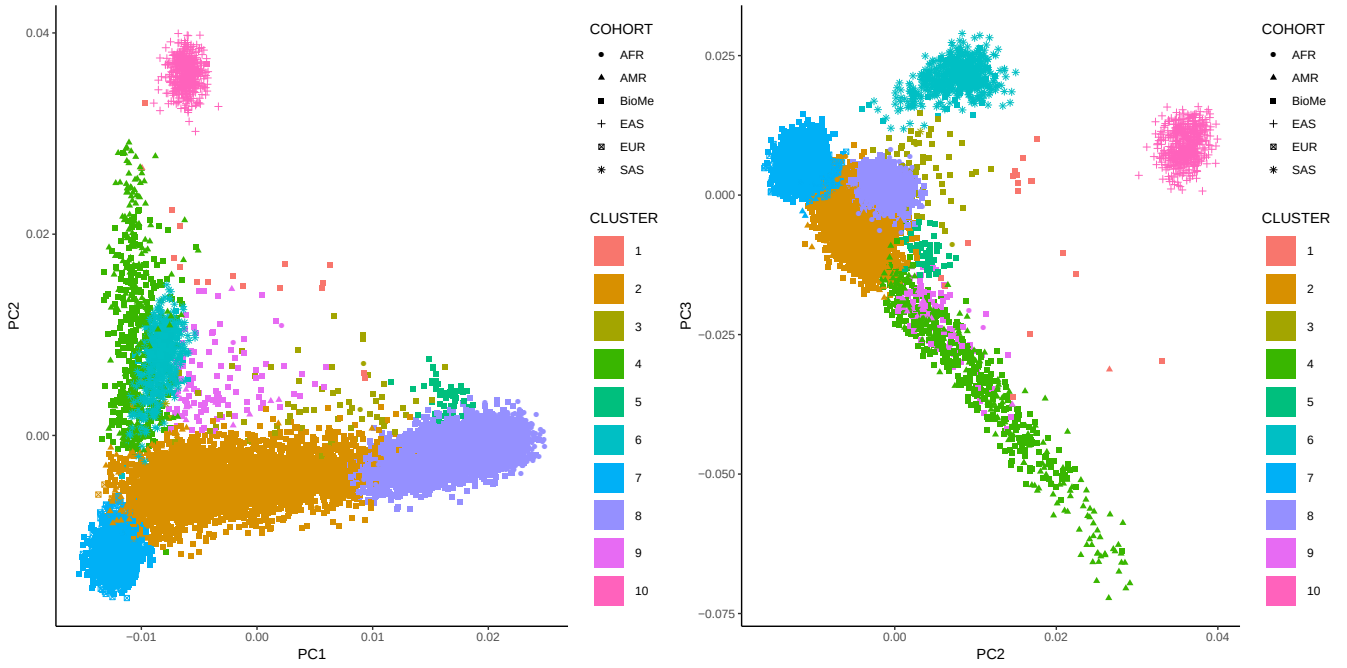
Using the principal components of ancestry as features, we employed the signal processing software Klustakwik [5] to model the array as a mixture of Gaussians, identifying clusters, or population groups/subgroups. In order to generate clusters of sufficient size for statistical association tests, we 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 6, 7, and 8 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. More information is available upon request



(a) PC1 vs. PC2

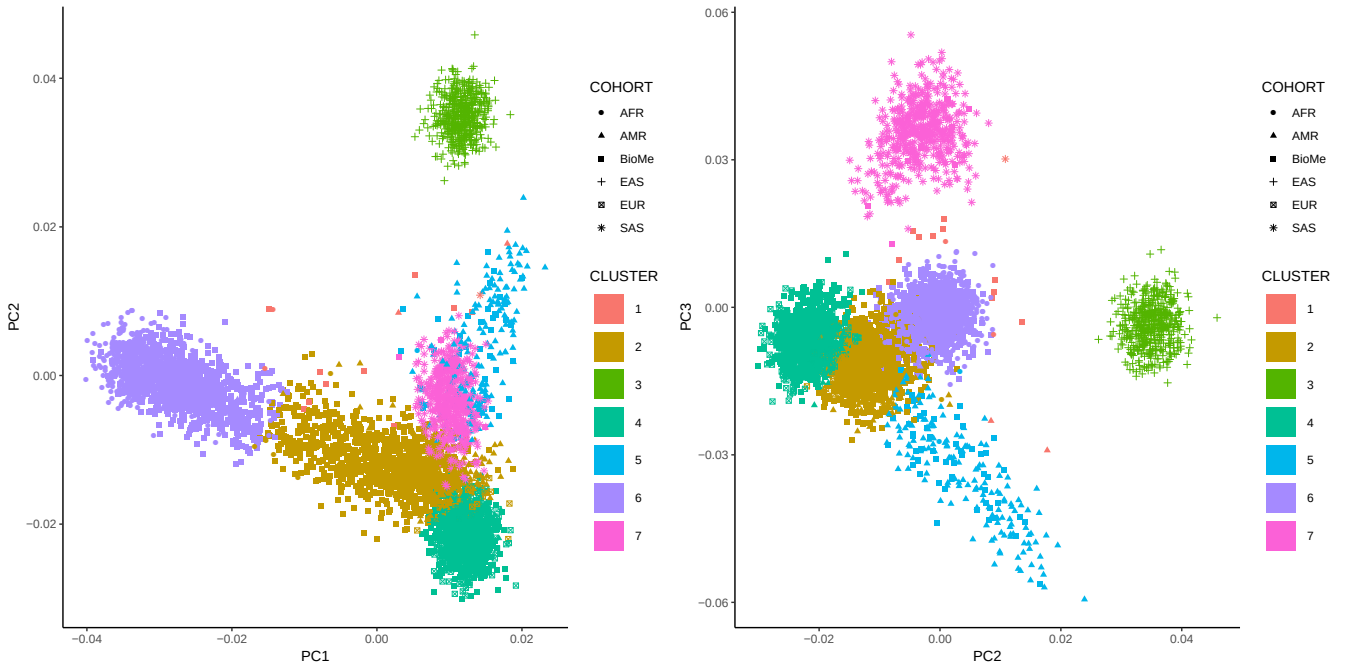
(b) PC2 vs. PC3

Figure 6: Population clusters for EX



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

Figure 7: Population clusters for ILL



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

Figure 8: Population clusters for AFFY

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 (EX > ILL > AFFY) and assigning each sample to the population determined using the highest technology.

Table 2: Inferred ancestry by dataset and cluster

	Population	Clusters	Samples
EX	AFR	2,5,9,10	5174
	AMR	3,4,7,12	5146
	EAS	6	2
	EUR	11	2379
	SAS	8	70
	Outliers	1	7
ILL	AFR	3,5,8	3783
	AMR	2,4,9	4187
	EAS	10	2
	EUR	7	1779
	SAS	6	57
	Outliers	1	17
AFFY	AFR	6	855
	AMR	2,5	1049
	EUR	4	596
	SAS	7	16
	Outliers	1	11

Table 3: Final inferred ancestry

Population	Samples
AFR	5205
AMR	5347
EAS	2
EUR	2407
SAS	72
Outliers	6

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 29,864 EX, 200,058 ILL and 193,393 AFFY variants remaining.

In order to identify duplicate pairs of samples, a filter was set to $Kinship > 0.4$. There were 3 EX and 2 AFFY 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, 1 EX and 2 AFFY samples have been reinstated. More information is available upon request

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. More information is available upon request

3.3 Sex Chromosome Check

Each array was checked for genotype / clinical data agreement for sex. There were 8 EX, 5 ILL and 3 AFFY 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. More information is available upon request

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.

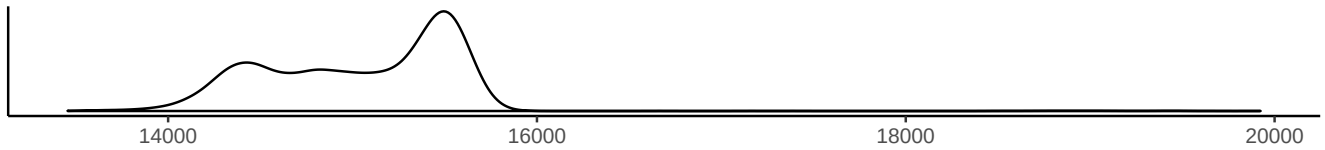
Table 4: Sample Metrics

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

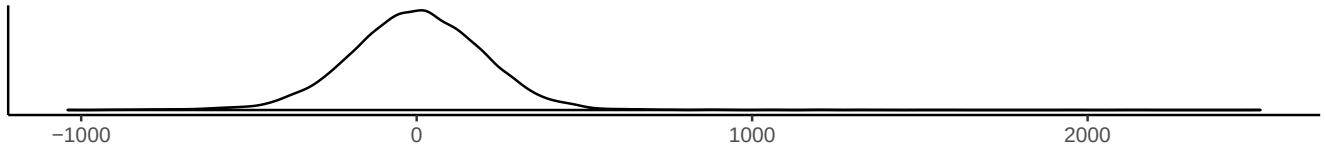
3.4.1 Principal Component Adjustment and Normalization of Sample Metrics

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 9 shows the n_non_ref metric for EX samples before and after adjustment.

Figure 9: 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 10, 11, and 12. 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

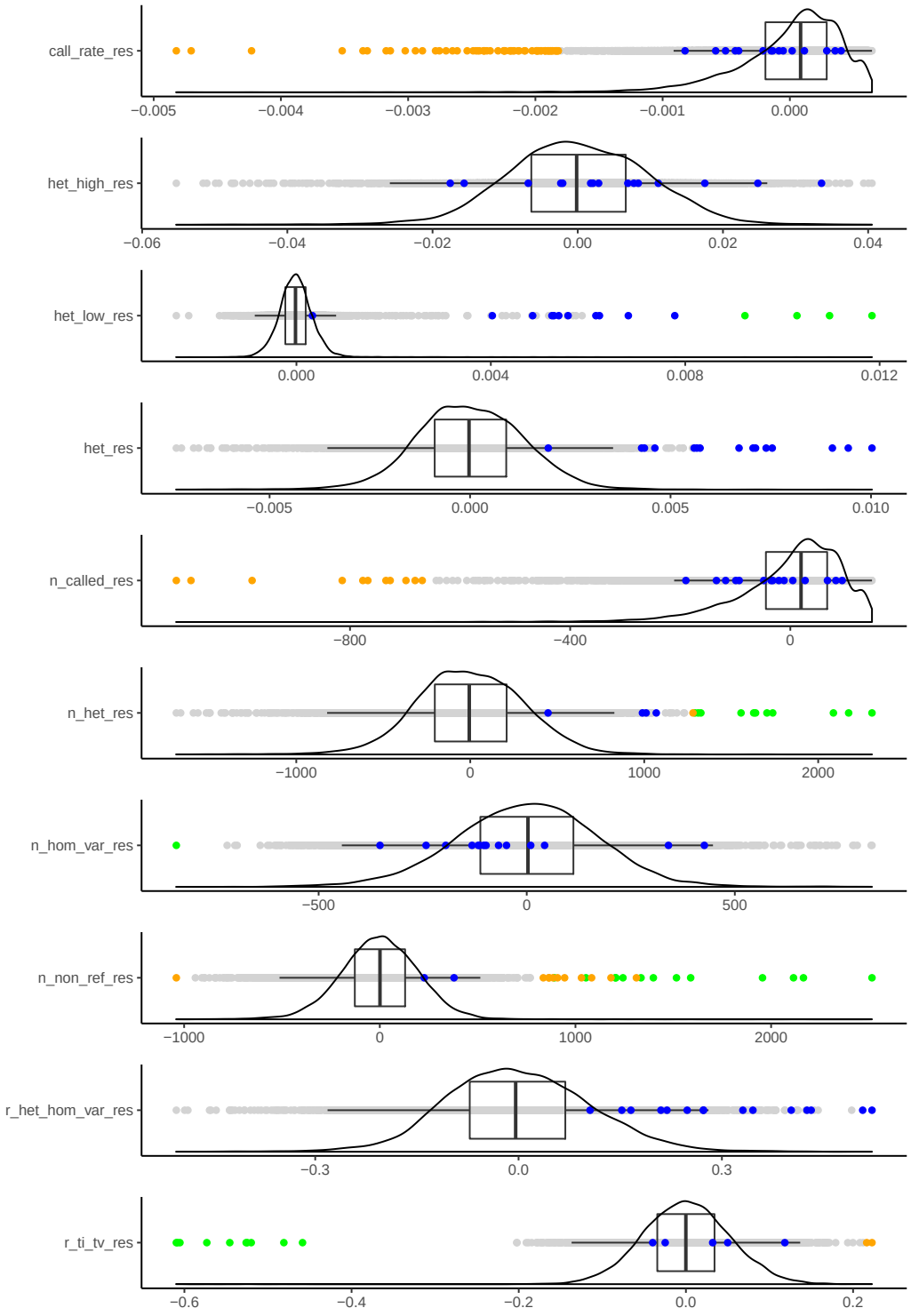


Figure 10: Adjusted sample metric distributions for EX

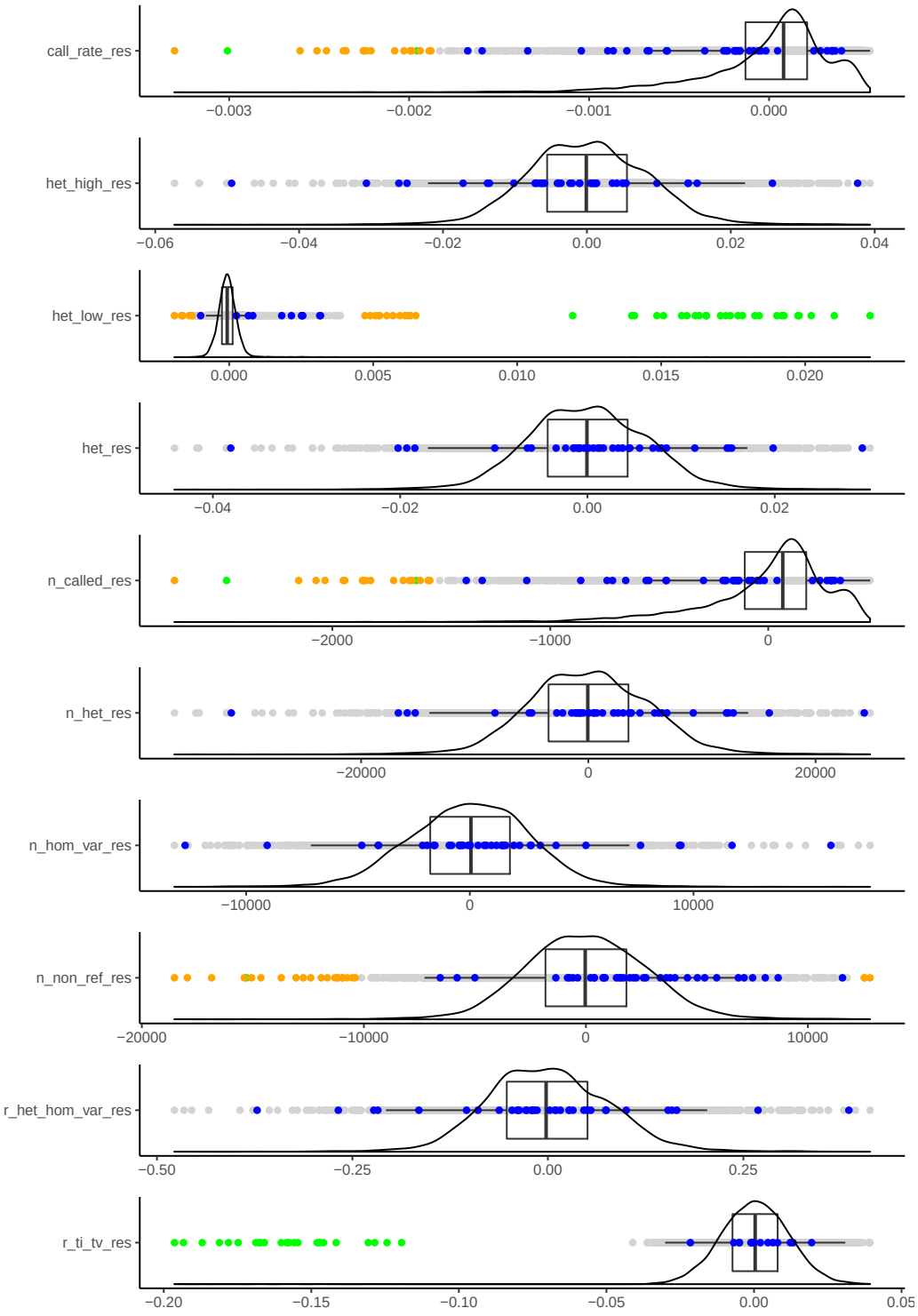


Figure 11: Adjusted sample metric distributions for ILL

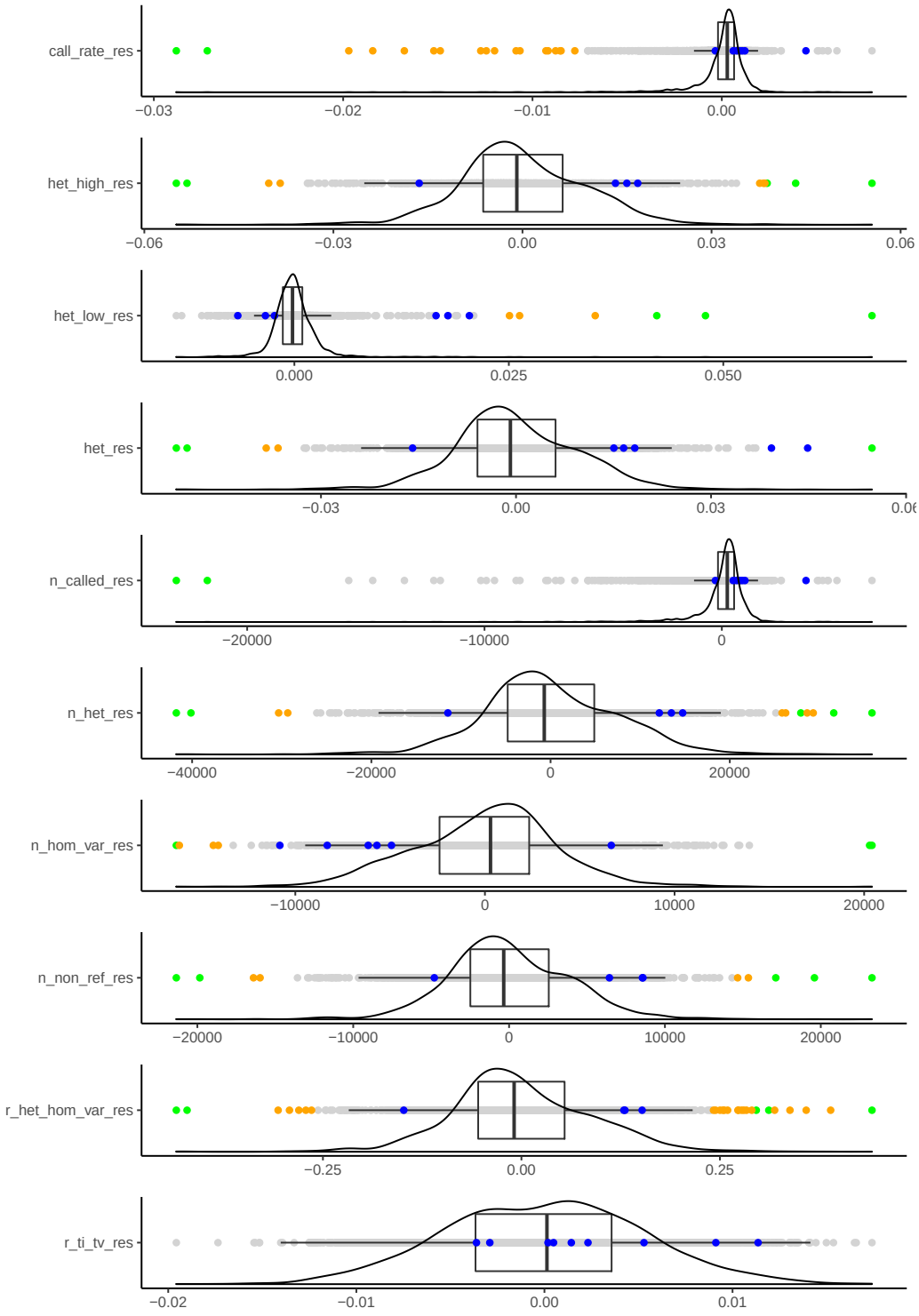


Figure 12: Adjusted sample metric distributions for AFFY

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 13 summarizes the samples remaining for analysis. Upon further inspection, no samples were manually reinstated during this step. More information is available upon request

Table 6: Samples flagged for removal

	AFFY	EX	ILL	Total
call_rate	25	80	55	124
het_high	14	15	38	54
het_low	12	15	56	69
het	11	15	38	51
n_called	9	26	55	69
n_het	15	16	38	56
n_hom_var	12	15	38	52
n_non_ref	13	26	61	87
r_het_hom_var	30	15	38	70
r_ti_tv	9	17	38	51
PCA(Metrics)	9	15	38	49
Metrics+PCA(Metrics)	45	92	92	181
Extreme Missingness	0	0	0	0
Duplicates	4	6	0	8
Cryptic Relatedness	0	0	0	0
Sexcheck	3	8	5	8
Ancestry Outlier	0	0	0	0
Total	58	112	103	203

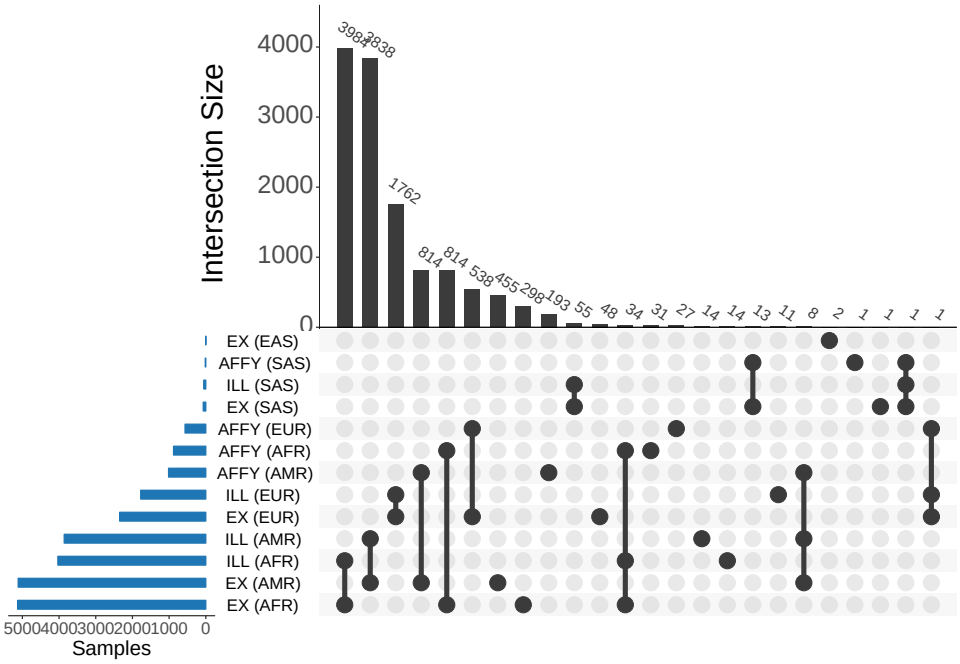


Figure 13: Samples remaining for analysis

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 301 EX, 579 ILL and 387 AFFY variants for removal. Figure 14 shows the number of variants remaining for analysis after applying filters.

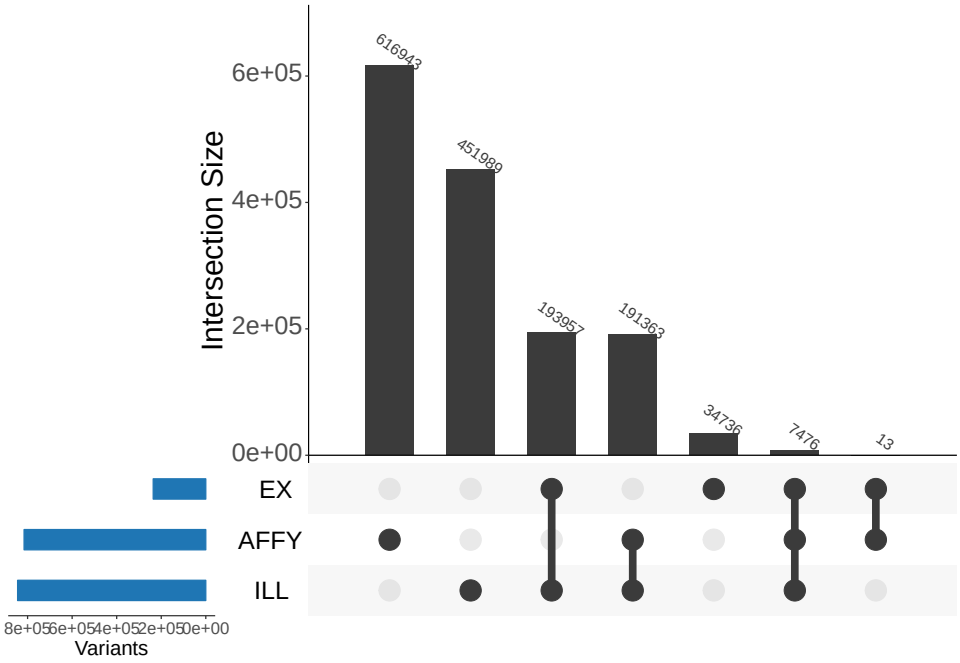


Figure 14: Variants remaining for analysis

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 (1.3-211-g92f442f) branch: master commit: 92f442fb6952fbc26e474fbf49b6db15a7370677 built on: 2019-05-03T16:01:23.778Z. 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-2.5.21. https://github.com/broadinstitute/dig-loam.