AMP-DCC Quality Control Report

CAMP

04/17/2019 (08:15)

Prepared by: Ryan Koesterer and Maria Costanzo on behalf of Broad Institute

Contact: AMP-DCC Analysis Team (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 CAMP dataset. We received genotypes for 3,732 unique samples genotyped on a single array. 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.

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.

This data consisted of a single genotype array (EX) which contained 3,732 remaining samples.

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	103117	2132	1352	0	39	0	74	728	106640
	(0,0.01)	2	133269	2606	194	0	231	0	89	1094	136302
	[0.01,0.03)	0	16496	491	38	0	68	0	2	197	17093
	[0.03,0.05)	0	8832	379	81	0	18	0	2	119	9310
	[0.05,0.10)	0	19311	674	34	1	26	0	1	292	20046
	[0.10,0.50]	0	236249	5678	279	0	15	0	2	4739	242221
	Total	2	517274	11960	1978	1	397	0	170	7169	531612

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.The resulting dataset consisted of 414,987 total variants.

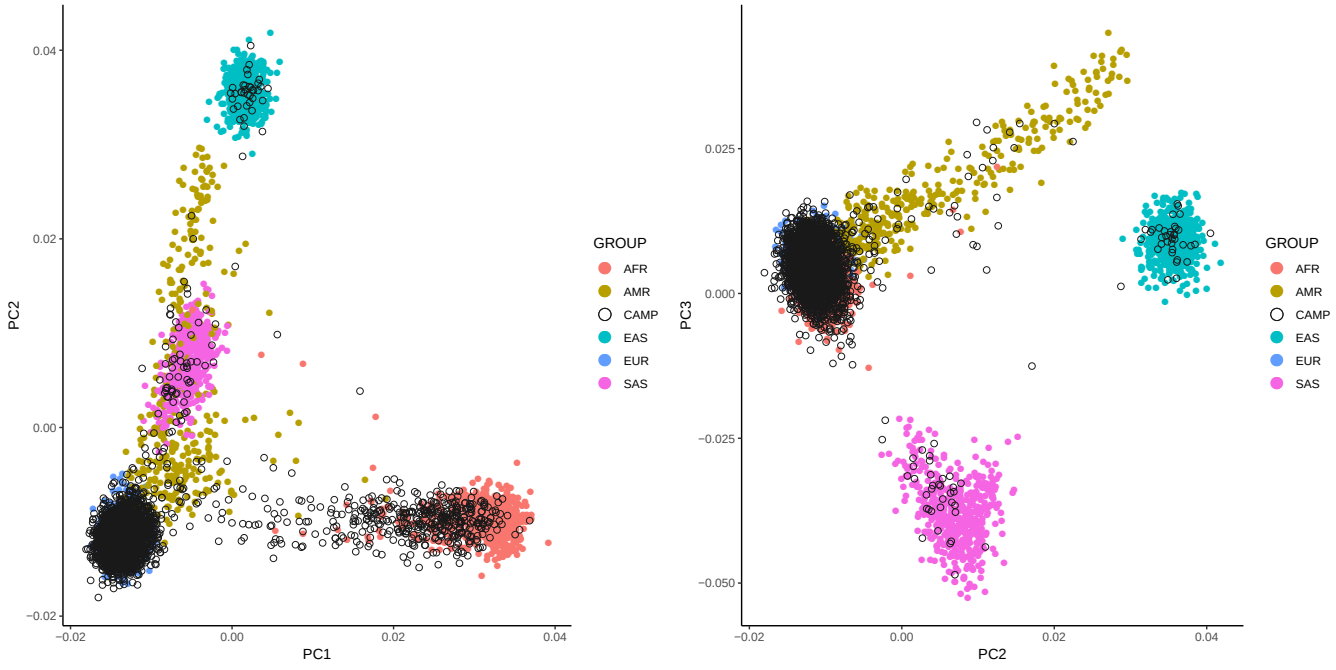
3 Sample QC

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, the 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 3,245 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.

Figure 1 displays plots of the top three principal components along with the 1000 Genomes major population groups.

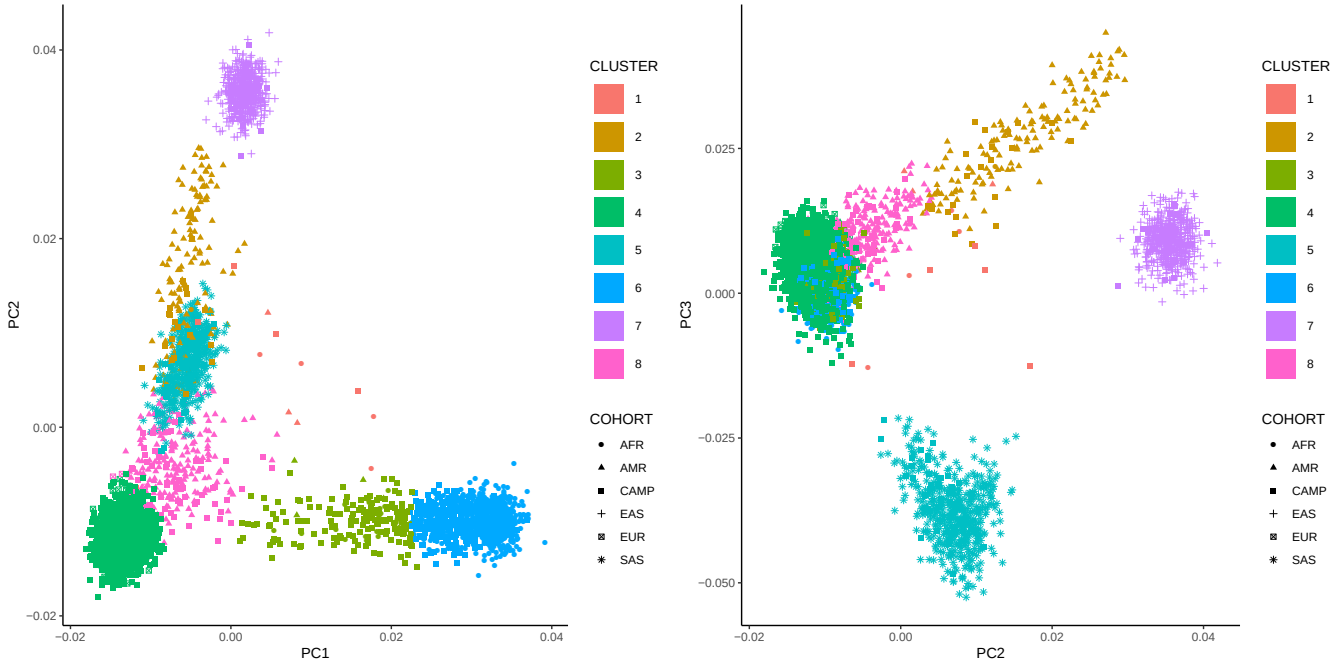


(a) PC1 vs. PC2

(b) PC2 vs. PC3

Figure 1: Principal components of ancestry for EX

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. Figure 2 clearly indicates 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 2: Population clusters for EX

The resulting clusters are then combined with the nearest 1000 Genomes cohort. Table 2 describes the classification using this method. Table eftable:ancestryFinalTable describes the final population assignments..

Table 2: Inferred ancestry by dataset and cluster

	Population	Clusters	Samples
EX	AFR	3,6	421
	AMR	2	21
	EAS	7	37
	EUR	4,8	3212
	SAS	5	36
	Outliers	1	5

Table 3: Final inferred ancestry

Population	Samples
AFR	421
AMR	21
EAS	37
EUR	3212
SAS	36
Outliers	5

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 83,710 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, both samples were removed. In this case, no 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 32 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.

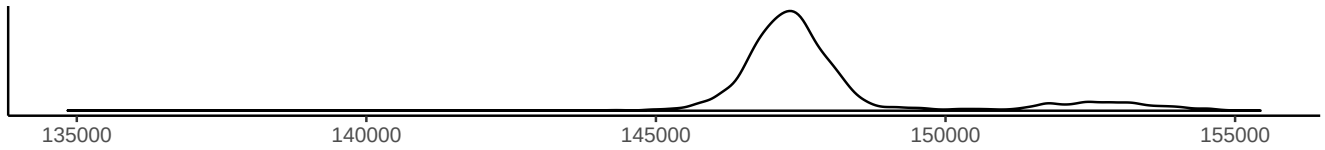
Table 4: Sample Metrics

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

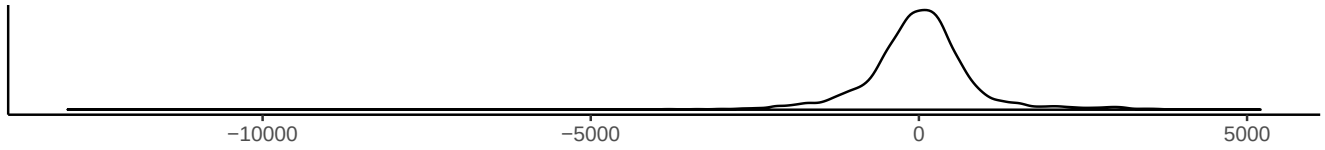
3.4.1 Principal Component Adjustment and Normalization of Sample Metrics

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

Figure 3: 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 all 10 PCARMs combined.

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 Figure 4. 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

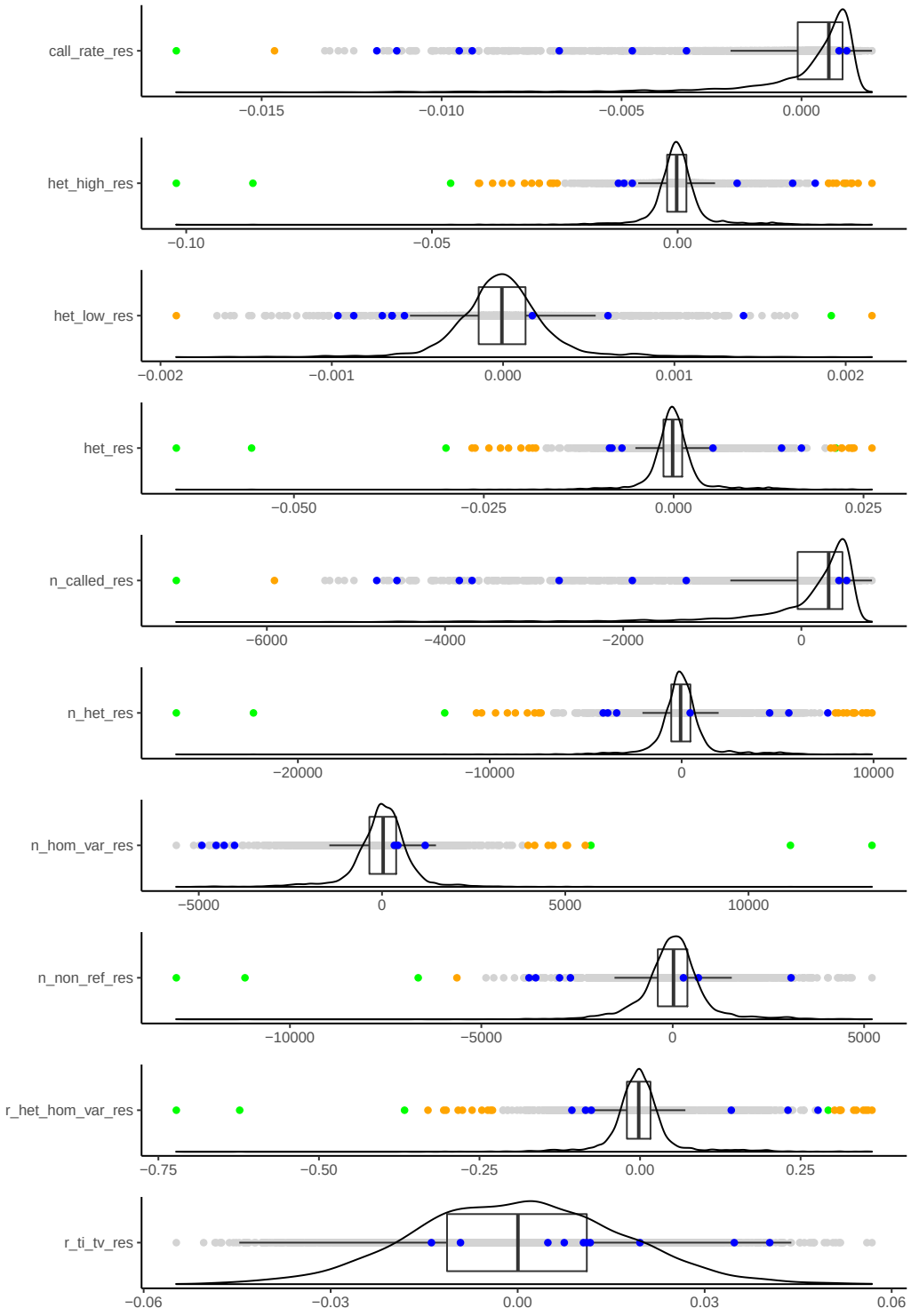


Figure 4: Adjusted sample metric distributions for EX

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 10 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 5 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

	EX	**Total**
call_rate	11	11
het_high	34	34
het_low	12	12
het	28	28
n_called	11	11
n_het	30	30
n_hom_var	17	17
n_non_ref	11	11
r_het_hom_var	30	30
r_ti_tv	10	10
PCA(Metrics)	10	10
Metrics+PCA(Metrics)	35	35
Extreme Missingness	0	0
Duplicates	0	0
Cryptic Relatedness	0	0
Sexcheck	32	32
Ancestry Outlier	0	0
Total	71	71

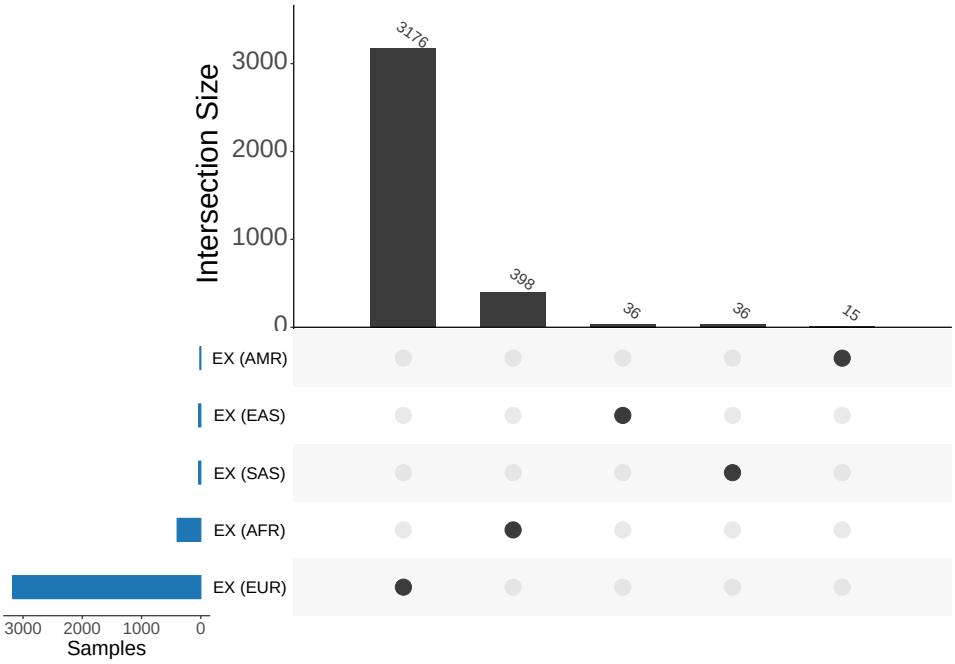


Figure 5: Samples remaining for analysis

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 10,134 variants for removal. After applying variant filters, there were 404,854 variants remaining for analysis.

5 Acknowledgements

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

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 (1.3-202-g7df4815) branch: master commit: 7df4815aa4da7f1845e3e40f931c11fda358c6af built on: 2019-04-10T18:38:22.667Z. 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-2.5.6. https://github.com/broadinstitute/dig-loam.