ALFA QC documentation
(Inclusion/exclusion criteria of study genotype datasets)

ALFA uses genotype datasets of studies released by dbGaP to calculate the allele frequencies. We check the datasets of each study and only include those that pass quality assurance (QC) checks for ALFA.

To find potential errors affecting genotype qualities in each study, we computed frequency from each study separately and QC the results using the same methods for the full build (See **Comparing AFs of AIMs)**. After excluding studies with potential errors reported by QC, we combined all passed studies to a full build to aggregate data across studies and compute allele frequency across studies.

For each ALFA full build, we compared the allele frequencies (AFs) with those calculated using data of 1000 Genomes Project (TGP). We compared AFs using both ancestry informative markers (AIMs) and all SNPs with RS IDs.

Comparing AFs of AIMs

For ALFA QC, we define an AIM as an SNPs with fixation index (F_{ST}), calculated either using the 1000 Genomes Project (TGP) data or ALFA build being checked, greater than a cutoff value. For each SNP in TGP, we calculate F_{ST} for differentiation of the six super-populations. For each SNP in the ALFA build, we calculate F_{ST} for differentiation of the ALFA populations. The minimum F_{ST} value is set to 0.2 for the current QC processes.

Assuming the two populations in each pair of the following table represent groups of subjects with similar ancestry backgrounds, we compared the AIM allele frequencies (AFs) for each population pair and plot them on scatter plots.

TGP super-population	ALFA population
EUR	EUR
AFR	AFO
EAS	EAS
AMR	LEN
SAS	SAS

For each AIM, the difference between the two AFs (Δ_{AF}) in each population pair is expected to be close to 0, since the subjects in the two populations have similar ancestry background. Defining AIMs with $|\Delta_{AF}| > 0.15$ as outliers, we checked the percentages and distributions of outliers and excluded a study from ALFA if any of the following statements is true for any population pair:

1) More than 0.1% AIMs are outliers.

2) Some unexpected distribution patterns, e.g., clusters, unexpected lines, are observed on the scatter plot by visual inspection. The expected distribution is that all AIMs are close to the $y=x$ line on the graph.

Comparing AFs of all SNPs

For ALFA builds created with multiple studies including multiple populations, the above QC check with AIMs can detect most of the errors. Suppose such a build contains multiple populations in a certain proportion, it is unlikely any of the study included in this build contains populations in the same proportion, and hence any systematic errors affecting the allele calling for any SNP will likely alter the population allele frequencies in a different proportion, which will make this SNPs an AIM as defined above, and make the error detectable using the above method.

However, if a build includes only one population, then no SNPs will be treated an AIM within this build. The above method will only check the AIMs based on the TGP data. If errors happen on SNPs other than these TGP-defined AIMs, the above method will not be able to detect them.

In order to detect errors in builds created using one or a few studies, we compared allele frequencies for all SNPs for the above mentioned five population pairs and find outliers. We marked an SNP as an outlier for a population pair if the absolute value of AF difference between TGP and ALFA build is greater than 0.3 and plot the outliers on a scatter plot for each population pair, excluding other SNPs since there are usually too many of them to be plotted. We calculated the percentages of the outliers and check the distributions of them by visual inspection. We excluded the build from ALFA if either of the following statements is true for any population pair:

1) More than 0.3% of SNPs are outliers.
2) The outliers are not randomly distributed, i.e., some patterns like clusters or lines are observed on the scatter plot.