Definitions and Explanations

Below is table with the definitions and explanations of thresholds and parameters used in this pipeline.

Parameter	Pipeline Argument	Definition
sample table[1]	-sampleTable	Sample information file in tab-delimited format extracted from genomeStudio or created by user. The following headers must be present (order does not matter, but headers are name and case-sensitive): • Sample ID • Call Rate • p10 GC
SNP table[1]	-snpTable	SNP information extracted from genomeStudio SNP table. The headers in this file should have the following information and Illumina filtering metrics with the following names (order does not matter, but headers are name and case-sensitive): • Chr • Name • Call Freq • Cluster Sep • AA T Mean • AA T Dev • BB T Mean • BB T Dev • AA R Mean • AB R Mean • BB R Mean
PLINK files[1,2]	-inputPLINK	PED file with associated MAP file or BED[2] file with associated BIM and FAM files

1 Required input for pipeline

2 binary PED (PLINK format) not to be confused with BED format adopted by UCSC for NGS DNA-,RNA-seq technologies

Parameter	Pipeline Argument	Definition
Genome Studio Final Report	--finalReport	This is an optional file. If provided and B Allele Frequency (BAF) as well as Log R Ratio (LRR) are provided, the mean calculation per sample for each of these metrics is output in a file called final_report_statistics_per_sample.txt. Note, this file should be sorted by sample name/IID and this needs to be the last column in the final report. Additionally, the following headers are required: • SNP Name • Log R Ratio • B Allele Freq • Sample Name (as last column in file) If the users chooses to use this parameter, it requires the parameter to be used as follows: final_report_full_path,header_line_number_minus_1 For example if my final report is called final and the header line starts on line 5, I would use the parameter as such: --finalReport /path/to/final,4
Array Type	--arrayType	The name of the chip or array used to plate SNPs
output directory	--outDir	Location and name of desired output directory to create the new file project folder. NOTE: all files created from PLINK in the pipeline will be stored where the input PLINK files are located. Only final PDF and output not generated from PLINK will be stored in this location. The default is the user's current working directory.
project Name	--projectName	The name of the project to be created in the output directory. All non-PLINK generated output will be stored in the desired output directory under the folder with this name. The project name must be unique and not already present. WILL NOT OVERRIDE EXISTING PROJECTS!
chip failure threshold	--chipFailure	[INT] *default: 1* This is the maximum total number of sex discrepancies OR sample missingness threshold failures PER CHIP before the chip is considered failing and a re-run or in-depth analysis needs to be performed.
sample call rate/frequency	--callrate	[FLOAT] *default: 0.991* The overall sample minimum call rate to be included in the sample set. Anything below this value will be removed from the sample space. This is essentially the ratio of SNPs that were called over all SNPs in a given sample.

Parameter	Pipeline Argument	Definition
snp call rate/frequency[3]	--snp_callrate	[FLOAT] *default: 0.97* The overall SNP minimum call rate in order to be included in the SNP sample space. Any SNP below this value will be removed. This is essentially the frequency, expressed as a ratio, at which the SNP was called across all samples. A low snp call rate/frequency can be indicative of a poor quality or failing SNP.
cluster separation[4]	--clusterSep	[FLOAT] *default: 0.30* Illimuna GenomeStudio metric that measures the separation in the AA, AB, BB clusters. Ranges from 0-1. Low cluster separation can be indicative of a poor quality or failing SNP. Any SNP below the threshold will be removed.
AA theta mean[4]	--AATmean	[FLOAT] *default: 0.30* Illumina GenomeStudio metric which measures the mean of the normalized homozygous AA theta values across all SNPs. Ranges from 0-1. A value above the default threshold may indicate that the AA cluster is far from the axis. Any SNP with a value above this threshold is removed from analysis.
AA theta standard deviation[4]	--AATdev	[FLOAT] *default: 0.06* Illumina GenomeStudio metrics which measures the standard deviation of the normalized homozygous AA theta values across all SNPs. Any SNP above this value is removed from analysis.
BB theta mean[4]	--BBTmean	[FLOAT] *default: 0.70* Illumina GenomeStudio metric which measures the mean of the normalized homozygous BB theta values across all SNPs. Ranges from 0-1. A value below the default threshold may indicate that the BB cluster is far from the axis. Any SNP with a value below this threshold is removed from analysis.
BB theta standard deviation[4]	--BBTdev	[FLOAT] *default: 0.06* Illumina GenomeStudio metrics which measures the standard deviation of the normalized homozygous BB theta values across all SNPs. Any SNP above this value is removed from analysis.

3 This threshold is only applied to autosomal chromosomes. X, Y, mitochondrial, and SNPs with an unknown or chr 0 specification are excluded from this filter.

4 Default value is set to Illumina recommended hard cut-offs

Parameter	Pipeline Argument	Definition
AA intensity mean[4]	--AARmean	[FLOAT] *default: 0.20* Illumina GenomeStudio metric to measure the mean normalized intensity of the AA homozygous cluster. Any SNP with an AA R mean equal to or below this threshold is removed due to unreliability resulting from low intensities.
AB intensity mean[4]	--ABRmean	[FLOAT] *default: 0.20* Illumina GenomeStudio metric to measure the mean normalized intensity of the AB heterozygous cluster. Any SNP with an AB R mean equal to or below this threshold is removed due to unreliability resulting from low intensities.
BB intensity mean[4]	--BBRmean	[FLOAT] *default: 0.20* Illumina GenomeStudio metric to measure the mean normalized intensity of the BB homozygous cluster. Any SNP with a BB R mean equal to or below this threshold is removed due to unreliability resulting from low intensities.
human genome build version	--genome_build	[select from b36-hg18, b37-hg19 or b38-hg38] *default: b37-hg19* The genome build that was used to map SNPs. This is used to determine where the X-chromosome boundaries are for sex-checking and imputation by PLINK.
estimated F coefficient for females	--maxFemale	[FLOAT] *default: 0.20* The maximum estimated F coefficient value in order for a sample to be imputed as female using the X-chromosome. This is the same default as suggested by PLINK, however, this should be adjusted accordingly.
estimated F coefficient for males	--minMale	[FLOAT] *default: 0.80* The minimum estimated F coefficient value in order for a sample to be imputed as male using the X-chromosome. This is the same default as suggested by PLINK, however, this should be adjusted accordingly.
p10GC	-	[FLOAT] This is a value calculated by Genome Studio of the 10th percentile of GenCall scores across all loci in a sample. It is generally positively correlated with overall sample call rate/frequency. Plotting this against the sample call rate can help determine which samples may have performed poorly or have low quality overall.

Parameter	Pipeline Argument	Definition
B Allele Frequency	-	[FLOAT] Sometimes abbreviated as BAF. This value can be found in the generated final report from Genome Studio. This is the ratio of samples that carry the B allele. A normal SNP would have a BAF 0f 0.0 for the AA locus, 0.50 for the AB locus, and 1.0 for the BB locus. If a SNP deviates from this, it may be indicative of a change in copy number variation.
Log R Ratio	-	[FLOAT] Sometimes abbreviated as LRR. This values can be found in the generated final report from Genome Studio. This is the logged ratio of observed probe intensity versus the expected probe intensity. Ideally this value should be 0 in a normal sample. Any major deviations from 0 may be indicative of a change in copy number variation.
SNP Name	-	This value can be found in the Genome Studio final report. If an rsID is available, this the associated rsID for a particular SNP. Otherwise, it is some other identifier or SNP name.
Sample ID	-	This value can be found in the Genome Studio final report. This is a unique sample identifier. The last part of the string contains the investigator's sample ID derived from the manifest file.
Allele 1 Forward/Allele2 Forward	-	This value can be found in the Genome Studio final report. This lists the genotypes reported in the forward orientation
Position	-	This value can be found in the Genome Studio final report. This contains the genomic position of the SNP in hg19.
Chromosome	-	Sometimes abbreviated as chr. This value can be found in the Genome Studio final report. This is the chromosome to which the SNP is mapped