**GenomeStudio Processing and Quality Control of the Genotype Data**

*Basic Protocol*

The protocol for the processing of raw Illumina genotype intensity data using GenomeStudio includes QC of genotype data using up-to-date criteria to identify and address problematic samples and SNPs. Those criteria and the QC steps to meet those criteria have been recommended by Illumina [42-45] and described by various authors [46-47]. Basically, after loading the raw data into GenomeStudio, the clustering of intensities for all SNPs is performed. The next step is to filter out low-quality samples. The best parameter to measure overall sample quality is the call rate, which measures the percentage of SNPs with genotype calls for a sample. The commonly used call rate standard by Illumina is ≥0.99. Illumina also recommends GenCall Score, which is a quality metric that indicates the reliability of the genotypes called, for samples across all loci (referred to as p10GC) to be 0.4–0.7. For SNPs, the most important QC parameter is the GenTrain score. The GenTrain score is computed from the GenTrain clustering algorithm. It is a measurement of SNP calling quality, ranging from 0 to 1, with higher value meaning better quality (Illumina recommends ≥0.4). The second most important QC parameter is the cluster separation score, which measures how well the AA, AB and BB clusters are separated. The cluster separation score also ranges from 0 to 1, with higher meaning better (more separation). The third most important QC parameter is call frequency, which measures the percentage of samples with successful calls for that SNP. The call frequency also ranges from 0 to 1, with higher meaning more samples have successful calls for this SNP (Illumina recommends ≥0.9). These three scores are often positively correlated, but they also identify unique scenarios to which only one of the three measures may be sensitive. Therefore, to determine whether manual re-clustering is needed, it is best to sort the SNPs by each of the three QC parameters, from small to large, and go through the SNPs with the lowest scores on any of the three measures.

*Pre-QC and Post-QC Results*

For the 55 samples, Call Rates and p10GC Scores ranged from 0.9625 to 0.9762 and from 0.3686 to 0.3718, respectively. For the HIHG control sample, Call Rate and p10GC Score were 0.9777 and 0.3716, respectively. For the 1,708,621 SNPs, Call Freq and GenTrain Scores ranged from 0.0000 to 1.0000 and from 0.0059 to 0.9655, respectively. The distribution of these SNPs per chromosome is presented below in Figure S1A.

There were a total of 31,395 SNPs that did not pass the No Call threshold of 0.15 (Call Freq = 0.0000). Among these SNPs, 927 were on Chr X, 1,999 on Chr Y, and 28,469 were on autosomes. The SNPs on autosomes included those that do not have a known Chr location, on Chr XY (pseudoautosomal SNPs), and on mitochondrial chromosome.

There were 1,677,226 SNPs, whose Call Freq ranged from 0.0179 to 1.0000. These SNPs were subjected to the QC steps [42-47]. From these SNPs, a total of 174,954 (10.43%) were removed at different steps of QC. Thus, a total of 206,349 (31,395 + 174,954) were removed. The distribution of these removed SNPs per chromosome is presented below in Figure S1B (count) and Figure S1C (percent). For the remaining 1,502,272 SNPs, Call Freq improved and ranged from 0.9107 to 1.0000. GenTrain Scores for these SNPs also improved and ranged from 0.3470 to 0.9655. Only 1,721 SNPs (0.001%) had GenTrain Score ≥0.3470 <0.4. The distribution of these remaining SNPs per chromosome is presented below in Figure S1D.

Post-QC, sample Call Rates and p10GC Scores improved and ranged from 0.9949 to 0.9997 and from 0.4134 to 0.4195, respectively. Similarly, for the HIHG control sample, Call Rate and p10GC Score improved to 0.9996 and 0.4196, respectively.



Figure S1. Pre- and post-quality control Infinium Multi-Ethnic Genotyping Array SNP distribution per chromosome.

A: Pre-quality control number of SNPs; B: Number of removed SNPs; C: Percent of removed SNPs; D: Post-quality control number of SNPs.

Chr = Chromosome

Chr 0 denotes SNPs that do not have a known Chr location.