Updated on 05/04/2019

Cytogenomic and Genotyping Data Analysis

The widespread use of microarrays allows gene expression profiling, genotyping, mutation detection, and gene discovery throughout the genome. This document aims to provide a workflow for analysis of Infinium® CytoSNP-850K v1.2 array data to identify genetic and structural variations.

The raw data will be analyzed using GenomeStudio® Genotyping Module and/or BlueFuse Multi software based on the reference human genome (hg38/GRCh38). After loading the raw data, the SNP manifest file (*.bpm), and standard cluster file (*.egt) into GenomeStudio, the clustering of intensities for all SNPs will be performed.

Genotypes are called by comparing customer-generated data with those in the standard cluster file. Genotyping calls for a specific DNA made by the calling algorithm (GenCall) which relies on information provided by the GenTrain clustering algorithm. The GenTrain score is a measurement of SNP calling quality, ranging from 0 to 1, with higher value meaning better quality. GenCall is more suitable to the identification of common SNPs. Cluster separation score measures how well the AA, AB and BB clusters are separated. Call frequency measures the percentage of samples with successful calls for that SNP.

Quality control analysis begins with preliminary sample quality evaluation to determine which samples may require reprocessing or removal. The best parameter to measure overall sample quality is the average call rate for each sample across all loci. Different genotyping arrays might have different call rate standards, yet the commonly used call rate standard is 95-98%. Any sample below the call rate standard will be excluded from further analysis. If manual re-clustering is needed, then we will sort the SNPs by each of the above three QC parameters (GenTrain score, Cluster separation score and Call frequency), from small to large, and will evaluate the SNPs with the lowest scores on any of the three measures. The quality assessment of SNPs on chromosome X and Y will be stratified by sex.

Detection of copy-number variants and chromosomal aberrations: The goal of the cnvPartition algorithm is to identify regions of the genome that are aberrant in copy number using two Infinium® assay outputs: the log R ratio (LRR) and B allele frequency (BAF). LRR is the log ratio of observed probe intensity to expected intensity, with any deviations from zero in this metric being evidence for copy number change. BAF is the proportion of hybridized samples that carry the B allele as designated by the Infinium assay. In a normal sample, discrete BAFs of 0.0, 0.5, and 1.0 are expected for each locus (representing AA, AB, and BB). Deviations from this expectation are indicative of aberrant copy number.

Additionally, we will export genotype data to PLINK format, which is the standard format for storing genotyping data, to perform PLINK level analysis.

Updated on 05/04/2019

REQUIREMENTS:

1. Intensity files (.idat). These files are the raw data files from the whole genome/exome chip. They should be provided by the facility that performed the genotyping.
2. Sample sheets. The sample sheets are CSV files that contain sample information, such as plate ID, cell ID, gender and so on. The sample sheets should be provided by the genotyping facility (See for a demo sample sheet at https://support.illumina.com/array/array_kits/cytosnp-850k_beadchip_kit/downloads.html)

DELIVERABLES:

1. Standard CNV Report—Lists each CNV and loss of heterozygosity (LOH) region for each selected sample.
2. Genome Browser screenshots of selected CNVs.
3. PLINK compatible input data

REFERENCES:

https://www.illumina.com/documents/products/technotes/technote_cnv_algorithms.pdf
https://support.illumina.com/array/array_kits/cytosnp-850k_beadchip_kit/downloads.html

Note: Emory Integrated Genomics Center (EIGC) offers high throughput SNP Genotyping using the Illumina NextSeq 550 (Cytogenomic array) or iScan (Genotyping array) platforms and any available Illumina (Infinium or Golden Gate® assay-based) genotyping arrays. Illumina also will design and produce custom Infinium chips, called iSelect, which can query between 3,000 and 200,000 SNPs per sample. The number of samples processed on each array varies by array type. For example, the HumanOmniExpress chip processes 24 samples while the Infinium® CytoSNP-850K v1.2 processes 8 samples per array. So be sure to check with EIGC (or Illumina web site) to determine the number of kits required to process your sample set. Contact the EIGC (EIGC@emory.edu) to discuss array selection and sample requirements.

Questions? Comments?

Please email us at EICC@emory.edu