CNV detection from genotyping arrays

Quality control pipeline

Lucia Trastulla, PhD student

Max Planck Institute for Psychiatry,
Ziller lab

1 Introduction

The quality control pipeline aims at the curation of the genotype calls from GenomeStudio output. The following pipeline is based on [1],[2] and is similar to what is described in here.

2 Steps

The first step is performed with the GenomeStudio software, when samples with a call rate < 0.98 are excluded and the SNPs statistic is recomputed, hence not including poor quality samples.
The following steps are performed in the Rscript QC_extract_LRR_BAF_run.R. All the parameters for the quality control described in the following paragraphs can be modified inside the R script. Chromosome 0 and MT (mitochondrial) are excluded from the analysis. The quality control is based on the result in SNP_table.txt and Full_Data_Table.txt

2.1 Quality Control of Autosomal SNPs

1. Chromosome from 1 to 22 and the pseudoautosomal region in chromosomes X and Y are considered.
2. SNPs with **Cluster Separation** ≤ 0.3 are excluded.
3. **Call Frequency** threshold is settled using the following function

$$f(n) = \begin{cases} 0.0004 \cdot n + 0.7804, & \text{if } 24 \leq n \leq 500 \\ 0.98, & \text{if } n > 500 \end{cases}$$

   where n refers to the number of samples in the study. The function grows linearly until the number of samples considered is 500 and then is constant. Thus, a study with a small number of samples is not excessively penalized in term of this parameter. SNPs with **Call Frequency** $\leq f(n)$ are excluded.
4. SNPs with **AB R Mean** ≤ 0.2 are excluded.
5. SNPs with **AB R Mean** ≤ 0.2 are excluded.
6. SNPs with **BB R Mean** ≤ 0.2 are excluded.

7. SNPs with **AB T Mean** ≤ 0.1 or **AB T Mean** > 0.9 are excluded.
8. SNPs with **Het Excess** < -0.9 or **Het Excess** > 0.9 are excluded.
9. SNPs with **MAF** > 0 and **AB Freq** $= 0$ are excluded.
10. SNPs with **AA Freq** $= 1$ and **AA T Mean** > 0.3 are excluded.
11. SNPs with **AA Freq** $= 1$ and **AA T Dev** > 0.06 are excluded.
12. SNPs with **BB Freq** $= 1$ and **BB T Mean** < 0.7 are excluded.
13. SNPs with **BB Freq** $= 1$ and **BB T Dev** > 0.06 are excluded.

2.2 Quality Control of Haploid SNPs

1. Chromosome X and Y without the correspondent pseudoautosomal region are considered.
2. The sex of each sample is inferred based on the following conditions:
   (a) **percentage** of genotype called as **AB** in chromosome X < 0.01
   (b) **percentage** of not called (**NC**) genotype in chromosome Y < 0.7

   If (a) and (b) are both satisfied then the gender is assigned as ”male”, if (a) and (b) are both not satisfied the gender is assigned as ”female”, otherwise is assigned as ”undefined”.
3. In **chromosome X**, exclude SNPs such that
$$\frac{\#\,\{\text{GT = AB in males}\}}{\#\,\text{males}} > 0.5$$
4. In **chromosome Y**, exclude SNPs such that
$$\frac{\#\{\text{GT} \neq \text{NC or R} > 0.2 \text{ in females}\}}{\#\,\text{females}} > 0.5$$
5. In **chromosome Y**, exclude SNPs such that **Call Frequency** for males $< f(\#\,\text{males})$ with f as defined in 2.1 (point 3)

3 Sample call rates

The new call rates for each sample is computed as the percentage of SNPs with a genotype call (hence different from NC). If the sample is classified as female, the chromosome Y (not PAR) is excluded from the computation.

References

[1] Illumina (2016). Infinium Genotyping Data Analysis. [Technical Data Sheet].
Online pdf

[2] Guo Y., He J., Zhao S., Wu H., Zhong X., Sheng Q., Samuels D.C., Shyr Y., Long J., Illumina human exome genotyping array clustering and quality control. *Nature protocols* (11) 2643-2662 (2014).