Chapter 3

Materials and Methods

This chapter describes the dataset and the SNP array platform used in Baependi Heart Study. The following sections will present the methods which are part of the pipeline we developed. The methodology is divided in two parts: CNV calling and CNV analysis.

The first part starts in Section 3.3, where we explain the procedures included in the pre-processing of genetic data, which main focus is to obtain two new values (Log R Ratio and B Allele Frequency) described in Section 3.4. Section 3.5 will define the Hidden Markov Model(HMM) used for CNV calling.

The second part begins with the pre-processing of the output from the HMM (Section 3.3) and ends with the description of the models used for infer the heritability of height and the CNV transmission rate (Section 3.7).

3.1 Dataset

Due to multiple waves of immigration, Brazil has a highly admixed population, which can drive to genetic and environmental influences on several traits. The Baependi Heart Study was created by the Heart Institute in 2005 to develop a longitudinal family-based cohort study to understand the variation of cardiovascular risk factors among the population and disentangle its genetic and environmental components. The study contains two steps of data collecting in accordance with a planned sample design, the first wave was performed between December 2005 and January 2006 and the second wave follow-up in 2010 (details are described in Egan *et al.* (2016)).

The data considered in this work is from the first wave of the described study and it provides information about 85 families (1,712 individuals), living in the village of Baependi, in the state of Minas Gerais, Brazil. Data from 631 nuclear families were available, with a size ranging from 1 to 14 offspring. The number of generations per family varied from 2 to 4 (54% of the families had 3 generations, and 45% had 2 generations). Only individuals aged 18 years or older were considered eligible for participating in the study. The mean age was 44 years, with a range of 18 to 100 years (de Oliveira *et al.*, 2008).

For each participant a questionnaire was used to obtain information regarding family relationships, demographic characteristics, medical history and environmental risk factors.

Anthropometric measures, physical examination and electrocardiogram of the participants were performed by trained medical students. Also, fasting blood glucose, total cholesterol, lipoprotein fractions and triglycerides were obtained by standard techniques in blood samples. Serum samples were stored at –80 ^{o} C and genomic DNA was extracted by standard procedures. From DNA samples, genotyping was made with Affymetrix Platform 6.0, 1120 CEL files were obtained.

3.1.1 SNP array platform

SNP array is a type of DNA microarray which is used to detect single nucleotide polymorphisms within a population. This platform 6.0 includes 906,600 SNPs markers and 946,000 CN

probes, in which 202,000 targets 5,677 CNV regions from the Toronto Database of Genomic Variant (Affymetrix, 2008).

As documented, the human genome of two individual are 99.9% identical at the nucleotide level and the presence of SNPs in the genome is the largest source of genetic diversity among humans, however, some regions of the genome can have no or few SNPs (Laframboise, 2009; Shen *et al.*, 2008); also, as described in 2.4, a single probe can represent multiple CNV regions. For these reasons, CN probes evenly spaced along the genome were added to the platform to include these regions that would be not covered by SNPs.

The procedure to obtain the intensity of the alleles for given marker using the Affymetrix assay is illustrated by Figure 3.1. For a given SNP, different oligonucleotides of 25 nucleotides (25-mer probes) for both alleles containing the SNP in different positions are used to bind to the DNA strand. When the probe is complementary to all 25 bases a brighter signal is detected than when there is a mismatching at the SNP site (Laframboise, 2009). The values of these intensities are stored in CEL files that will be used for the CNV analysis.

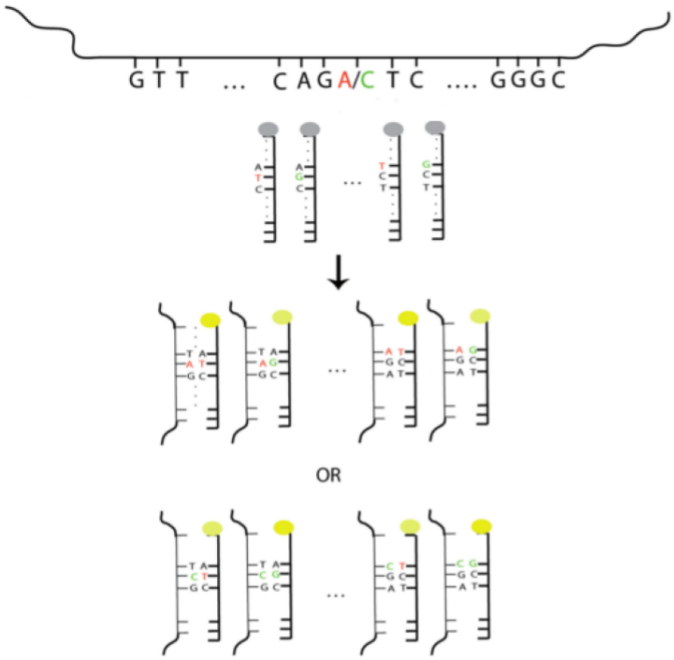


Figure 3.1: *Illustration of signal extraction for a given marker. Source: Laframboise (2009).*

3.2 Methodology Overview

The methodology can be summarized by Figures 3.2, 3.3 and 3.4, which describe the pre-processing of SNP data, the CNV calling and the CNV analysis, respectively.

For the pre-processing of SNP data and the CNV calling, the softwares *Affymetrix Power Tools*, *PennCNV* and libraries from R package were used. Briefly, the CEL files from Affymetrix 6.0 platform are used to obtain the intensities of alleles A and B for each single nucleotide polymorphism (SNP). Based on these values, the genotypes (AA, AB e BB) are estimated by clustering algorithms. Then, using the intensities values, new values were obtained by polar transformation: log R ratio (LRR) and B Allele frequency (BAF). These new information is used in a hidden Markov model for CNV estimation.

The following functions were used in the pre-processing described in Figure 3.2 (Wang *et al.*, 2007a):

1. Function: *apt-probeset-genotype* from Affymetrix Power Tools.

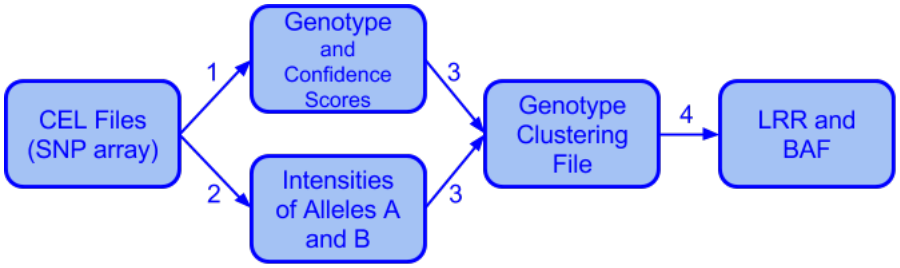


Figure 3.2: *Flowchart of CNV Calling. The number indicates which function was used*

The signal intensity values for PM probes will be normalized using quantile normalization (Section 3.3.1). Then, the function will apply the median polish (Section 3.3.2) to get the final intensity values for alleles A and B for each SNP.

2. Function: *apt-probeset-summarize* from Affymetrix Power Tools.
   Given the CEL files, this function will generate the genotype calls using the Birdseed algorithm. For each SNP in each sample, the genotype will be coded as 0, 1, 2 or -1 for AA, AB, BB and missing values, respectively, with its corresponding confidence scores. Also, a final report will infer the sample sex.

3. Function: *generate_affy_geno_cluster.pl* from PennCNV.
   This function will generate canonical genotype clustering file, based on the output files from functions 1 and 2.

4. Function: *normalize_affy_geno_cluster.pl* from PennCNV.
   The calculation of LRR and BAF for each SNP are made using the genotype clustering file and intensities values of the alleles A and B, more details are in Section 3.4.

The pre-processing phase generates auxiliary files, such as the genotype clustering file and genotype confidence scores, and the file containing the LRR and BAF for all markers. These information is used to the CNV calling (Figure 3.3). The following functions were used in this process (Wang *et al.*, 2007a):

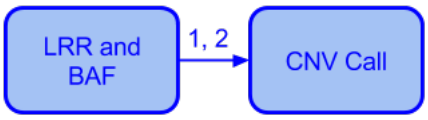


Figure 3.3: *Flowchart of CNV Calling. The number indicates which function was used*

1. Function: *kcolumn.pl* from PennCNV.
   This function splits the output from function 4 based on sample. Once, the HMM performs the CNV calling per sample.

2. Function: *detect_cnv.pl* from PennCNV.
   The CNV calling is performed for each sample (Section 3.5). Some additional information from HapMap are used, for example, the SNP coordinates and the population frequency of B allele.

PennCNV is used once it (1) estimates locus-level copy number, (2) performs segmentation, (3) evaluates CNV-specific quality-control metrics within a single software package, (4) has relatively small bias and variability and (5) detects regions while maintaining an estimated false-positive rate (Eckel-Passow *et al.*, 2011).

The found CNV regions are specific for each sample (individual), so the first filter is based on the quality control of the sample CEL file. Then, it is built a new set of minimal CNV regions, defined by the regions overlaps of all samples. A second filter is made, removing all minimal regions with no ou few copy number variations. The final regions are then ready for the CNV analysis of this project.

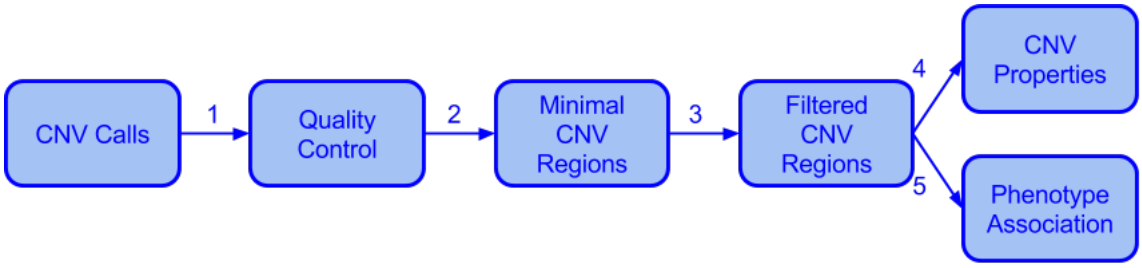


Figure 3.4: *Flowchart of CNV Analysis. The number indicates which function was used*

The following functions were used in the process of CNV analysis described in Figure 3.4:

1. Function: *filter_ cnv.pl* from PennCNV.
   PennCNV do not filter the samples, but this function extracts from the summary files the values of mean, median and standard deviation of LRR and BAF for each sample. With this information, we can evaluate the quality of the CNV calling based on our own criteria.

   The output from *filter_ cnv.pl* are loaded in R to select the best samples. Then, defined filters described in Section 3.6.1 are applied to this file, and a list of samples that passed the criteria is obtained.

2. Function: *CNTools* package from Bioconductor (Zhang (2017)).
   Each sample contains its own CNV regions. However, for posterior analysis, we need the same variables for all samples. The solution is to identify the minimal regions (Section 3.6.2), which are the overlapping of all identified regions. *CNTools* package was created with a similar aim, so we made some adaptations for our necessity.

3. Function: Basic functions from R.
   The obtained minimal regions contains regions that all samples are normal (contains two copies) or only a few samples has a mutation. For this reason, we filtered the regions so that at least 5% of the samples has an altered number of copies using the basic functions from R.

4. Function: Basic functions from R and *polygenic* from Solar (Blangero *et al.*).
   The script to obtain the characteristics of the CNV include simple functions from R. To estimate the CNV transmission rate, we use the *polygenic* function in R.

5. Function: *polygenic* from Solar and *kinship2* package from R (Therneau e Sinnwell (2015)).
   To estimate the heritability of phenotypes using CNVs as covariates, we use the *polygenic* function from Solar and the function *lmkin* from *kinship2* package.

3.3 Pre-processing of SNP data

The CEL file output is a specific file from Affymetrix. They store the intensity values of each probe array and its standard deviation, a flag to indicate an outlier, a user defined flags, and the number of pixels values collected from an Affymetrix GeneArray scanner (Affymetrix, 2009a,b).

To generate the intensity values and genotype calls from the CEL files obtained from our sample, the *The Affymetrix Power Tools (APT)* was used. The procedure described in McCall *et al.* (2010) involves:

1. Quantile normalization;
2. Median polish;
3. Genotype Calling (Birdseed).

The methods 1 and 2 are used to extract the intensity values of alleles A and B. Method 3 is used for the genotype calling.

3.3.1 Quantile normalization

The quantile normalization is a necessary procedure to remove variation due to target preparation and hybridization (McCall *et al.*, 2010). The step aims to make the distribution of probe intensities for each array in a set of arrays more comparable. This method is usually used in biostatistics for multiarray analysis.

The algorithm for this procedure is described in (Bolstad *et al.*, 2003) and it was developed based on the fact that two data vectors with the same distribution will show a straight diagonal line across the origin given by the unit vector $(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ in a quantile-quantile plot.

This can be generalized for n data vectors with the same distribution, which n-dimensions quantile-quantile plot will be a straight line across the origin given by the unit vector $(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$.

Thus, the quantile normalization is the procedure of projecting the points of a n-dimensional quantile plot onto the diagonal, following the 5-steps algorithm:

1. Given n arrays of length p, create the X matrix of dimension $p \times n$ which each array is a column;
2. Sort each column of X to give X_{sort};
3. Take the means across rows of X_{sort};
4. Substitute each element in the row by the row mean to get X'_{sort};
5. Get $X_{normalized}$ by rearranging each column of X'_{sort} to have the same ordering as original X.

Even though, there is the quantile function for R (Package 'preprocessCore) (Bolstad, 2001), a simplified example of how quantile normalization works using a R script can be found in Appendix 4.3.5.

3.3.2 Median polish

Based on the robust multiarray analysis (RMA) (McCall *et al.*, 2010), after the quantile normalization, the median polish is performed for removing possible outliers in the array. This technique described in Tukey (1970) extracts the row and column effects in a two-way table using medians and it is similar to the ANOVA, using median instead of mean.

For a given $X_{p\times n}$, x_{ij}, for $i = 1, \ldots, p$ and $j = 1, \ldots, n$, can be decomposed as in Equation 3.1, in which, t is a constant, r_i is the effect associate to the $i-th$ row, c_j is the effect associate to the $j-th$ column and res_{ij} is the residual associated to each element of the matrix.

$$x_{ij} = t + r_i + c_j + res_{ij} \tag{3.1}$$

The median polish follows the algorithm:

1. Given a matrix $X_{p\times n}$, set the values $t = 0$ as the constant, r and c as a vector of zeros with length p and n to be the column and row effects, respectively, and $\delta = 0$ as auxiliary variable;
2. For $X_{i.}$ $(i = 1, \ldots, p)$, compute the median (m_i), subtract m_i for each element of $X_{i.}$ and sum m_i to r_i;

3. δ is defined as the median of c. Subtract δ from c. Then, sum up δ and t;

4. For $X_{.j}$ $(i = j, \ldots, n)$, compute the median (m_j), subtract m_j for each element of $X_{.j}$ and sum m_j to c_j. This new X represents the residual matrix;

5. δ is defined as the median of r. Subtract δ from r. Then, sum up δ and t;

6. If the sum of the absolute values of X is equal to 0 or X is very similar to the previous X, the values of the constant, row and column effects and residual are inferred. Otherwise, repeat the steps 2-5 until one of the criteria is set.

As this method is used to remove the outliers, the residual matrix is subtracted from the original X. This means that the outlier will continue in the dataset, but the residual associated to it will the removed.

This technique is also used for estimating the expression of gene given the normalized expressions of a set of probes within the gene region. In this case, the value t is used (McCall *et al.*, 2010).

3.3.3 Genotype Calling

Birdseed is a genotyping and clustering algorithm for Affimetrix SNP arrays platforms. It uses a customized Expectation-Maximization (EM) algorithm to fit two-dimensional Gaussians to normalized and summarized (median polish) SNP data (A-signal vs. B-signal), producing clusters, which identify genotypes, and confidence scores for every individual at each SNP (Broad Institute, 2008).

The algorithm described by Korn *et al.* (2008)(Supplemental Material) has the following steps for each SNP:

1. The initial conditions for each cluster are based off a prior models file that contains SNP-specific estimates of cluster locations and variances learned from samples of known genotype;

2. The prior model is scaled by a value s to be in the same intensity space as the samples. This value s is defined as:

$$s = \frac{n}{d} = \frac{\frac{\sum_i \sqrt{I_{ai}^2 + I_{bi}^2}}{N}}{\frac{\sum_{c=AA,AB,BB} (W_c + 0.1)\sqrt{\mu_{ca}^2 + \mu_{cb}^2}}{\sum_{c=AA,AB,BB} (W_c + 0.1)}} \tag{3.2}$$

   which,

   n = scale for each SNP as the average distance of a sample from the origin;
   d = weighted average of the prior model means from the origin;
   I_i = Intensity of SNP i for allele A or B;
   N = Number of samples;
   μ_c = mean of cluster c;
   W_c = Expected weight of cluster c.

3. The data is fitted in four different 2d Gaussian Mixture Models (GMM) which are based on the number of possible clusters (one to three). The initial conditions of each model can be found on Korn *et al.* (2008), and models 3 and 4 differ only by these values, once they both consider the number of clusters equals to three.

   In general, this step performs a modified Expectation-Maximization (EM) algorithm, which includes two steps: Expectation, where calculates the probability of each sample to belong to each cluster, and Maximization, where updates the parameters of each cluster based on the results of the Expectation step. This step repeats until convergence or the maximum of iterations is reached.

4. Model selection is then performed following different criterias, such as BIC information criterion, closeness between final means (μ_1, μ_2, μ_3) and expected means (μ_{AA}, μ_{AB}, μ_{BB}).

5. Once the model is selected, genotype and confidence score are calculated.

The crlmn algorithm (Scharpf *et al.*, 2011; Wang *et al.*, 2008b) is also an alternative procedure for genotype calling.

3.4 Log R Ratio (LRR) and B Allele Frequency (BAF)

Log R Ratio and B Allele frequency are the parameters used by PennCNV and other methods for inferring CNVs. They are preferred instead of the normalized intensity values for simplicity, once they are based on polar coordinates and the intensities values are likely to create clusters. These parameters also take into account a reference value, making them comparable to HapMap population.

Therefore, a polar coordinate transformation of two-channels (two alleles, A and B) normalized intensity data is performed for each SNP, obtaining a intensity value, called R, and an allelic intensity ratio, called θ (Peiffer, 2006). Geometrically, R is the distance of the dot to the origin (Equation 3.3) and θ is the angular coordinate (Equation 3.4).

$$R = A + B \tag{3.3}$$

$$\theta = \frac{arctan(A/B)}{\pi/2} \tag{3.4}$$

3.4.1 Log R Ratio (LRR)

The log R ratio (or log2 ratio, or LR, or LRR) is the most common normalization and can be defined as:

$$LRR = log_2 \frac{R_{observed\ intensity}}{R_{reference\ intensity}} \tag{3.5}$$

The LRR equation shows that it is critical choosing a proper reference panel as it can affect all subsequent analyses (Gold Helix, 2014). In this case, the control panel used is HapMap. Based on Equation 3.5, it is easy to see that when the $R_{observed\ intensity}$ is equal to $R_{reference\ intensity}$, LLR is 0 and no variation in the number on copies is detected, a possible deletion can be seen when $LRR < 0$ and a duplication when $LRR > 0$.

3.4.2 B Allele Frequency (BAF)

The B allele frequency (BAF) is calculated based in the relative allelic signal intensity ratio explained by the equation 3.4, in which, A and B are the normalized intensity of alleles A and B. Equation 3.6 shows the values of BAF dependent of θ, where the values of θ_{AA}, θ_{AB} and θ_{BB} are the values of θ for three canonical genotype clusters generated from a large set of reference samples (Wang *et al.*, 2007b).

$$BAF = \begin{cases} 0 & \text{if } \theta < \theta_{AA} \\ 0.5\frac{(\theta-\theta_{AA})}{(\theta_{AB}-\theta_{AA})} & \text{if } \theta_{AA} \leq \theta < \theta_{AB} \\ 0.5 + 0.5\frac{(\theta-\theta_{AB})}{(\theta_{BB}-\theta_{AB})} & \text{if } \theta_{AB} \leq \theta < \theta_{BB} \\ 1 & \text{if } \theta \geq \theta_{AA} \end{cases} \tag{3.6}$$

The BAF is an additional information used to infer the CNV. It can be interpreted as the presence of the genotype has a B allele, then $BAF = 0$ indicates genotypes as A/A and A/-,

$BAF = 0.5$ as A/B and $BAF = 1$ as B/B and B/-. Complex phenotypes such as AA/B and AA/AA can alter the value of BAF and for this it has a bigger per-probe signal-to-noise ratio (SNR) than the LRR (Alkan *et al.*, 2011).

On the bright side, using the BAF allows the detection of events of neuter copy, as the uniparental dissomy (UPD), in which the prole receives two copies from one parent and no copy from the other parent, and the identity by descent (IBD), in which the segment is replaced by the other allele(Alkan *et al.*, 2011).

Figure 3.5 shows how the combination of LRR and BAF can indicate the number of copies, including the case of UPD and IBD, where LRR is two as expected, but BAF indicates the lack of heterogeneity.

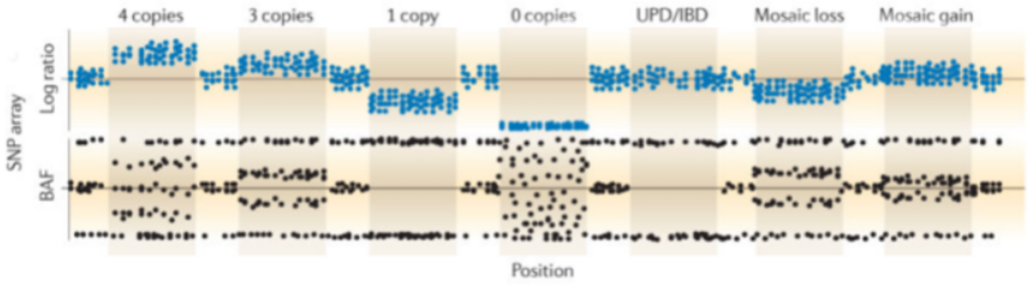


Figure 3.5: *Values of LRR and BAF for each case of variation in the number of copies, including more complex cases when the number of copies is equal to two even though a mutation is present. Source: Alkan et al. (2011).*

3.5 Hidden Markov Models

A Markov process considers that the probability of observing a particular state at a particular time point only depends on the state at the previous time point. A Hidden Markov model (HMM) is a sequence model that assumes the sequence follows a Markov process with unobserved states, dealing with 'labeling' problems, such as gene identification (label nucleotides as exons, introns, or intergenic sequence) and CNV detection (label markers as deletion, normal, or duplication) (Eddy, 2004).

In Wang *et al.* (2007b), a first order HMM was developed for CNV calling, once two adjacent SNPs are more likely to be found at the same state when they are close linkage.

For each SNP, let $\{r, b, z\}$ denote the LRR, BAF and the copy number state, respectively.

For a HMM, it is necessary some definitions (Eddy, 2004):

- A symbol and the number of symbols, K: The model goes through a sequence of markers (SNP and CN probes) and there are two continuous values associated to for each marker: The LRR can vary from -2 to 2 and the BAF, from 0 to 1, thus, $K = \infty$;
- Number of states, Z: In this case, they are six states as described in Table 3.1.
- Emission probability: $e_z(x)$ for each state z. This is the probability of r and b given the state z. It sums to one over K symbols k, $\int_k e_z(k) = 1$;

  Since we have two different parameters, there are two different emission probabilities, one for r (Eq. 3.7) and another for b, given by Eq. 3.8 (Wang *et al.*, 2007b).

  – LRR:

Copy state (z)	Total Copy	Description (for autosomes)	CNV Genotype
1	0	Deletion of 2 copies	-
2	1	Deletion of 1 copy	A, B
3	2	Normal State	AA, AB, BB
4	2	Normal State (LOH)	AA, BB
5	3	Single Copy duplication	AAA, AAB, ABB, BBBB
6	4	Double Copy duplication	AAAA, AAAB, AABB, ABBB, BBBB

Table 3.1: *Each state has a different distribution of CNV genotypes*

$$P(r|z) = \pi_r + (1-\pi)\phi(r;\mu_{r,z},s_{r,z}) \tag{3.7}$$

Where P is the probability of a LRR(r) given the state z, ϕ is the density function of a normal distribution with mean $\mu_{r,z}$ and standard deviation $s_{r,z}$ and π is a uniform distribution for correcting possible random fluctuations.

– BAF:

$$\begin{aligned} P(b|z) = & \underbrace{\pi_b + (1-\pi_b)\sum_{g=2}^{K(z)-1} BN[g-1;K(z)-1,p_B]\phi(b;\mu_{b,g},s_{b,g})}_{\text{A}} \\ & + \underbrace{(1-\pi_b)BN[0;K(z)-1,p_B][\mathbb{1}_{b=0}M_0 + \mathbb{1}_{0<b<1}\phi(b;\mu_{b,1},s_{b,1})]}_{\text{B}} \\ & + \underbrace{(1-\pi_b)BN[K(z)-1;K(z)-1,p_B][\mathbb{1}_{b=1}M_1 + \mathbb{1}_{0<b<1}\phi(b;\mu_{b,K(z)},s_{b,K(z)})]}_{\text{C}} \end{aligned} \tag{3.8}$$

K(z) is the number of possible genotypes for copy number at state z. $BN[g-1;K(z)-1,p_B]$ indicates the frequency for a genotype with g copies of allele B and p_B is the population frequency of B allele.
As given in Equation 3.6, BAF assumes three different values (0, 0.5, 1). For the BAF emission probability ($P(b|z)$), the terms A, B and C add values based on three cases: For $0<b<1$, its distribution is modeled as a normal mixture (from terms A, B and C); for BAF=0 (terms A and B) and BAF=1 (terms B and C), its distribution is modeled by a mixture of point mass at $0(M_0)$ or 1 (M_0) and a truncated normal (Wang *et al.*, 2007b).

- Transition probabilities $t_i(j)$: This is the probability of going from state i to a state j (including itself). It sums to one over the Z states j, $\sum_j t_i(j) = 1$.

In this case, it means the probability of having a copy number changing between two adjacent SNPs. For this, PennCNV uses the following Equation 3.9:

$$P(z_i = l|z_{i-1} = j) = \begin{cases} 1 - \sum_{k=2}^{6} p_{j,k-1}(1 - e^{\frac{-d_i}{D}}) & \text{if } l = j \\ p_{j,j-1}(1 - e^{\frac{-d_i}{D}}) & \text{if } l \neq j \end{cases} \tag{3.9}$$

where, D is a constant that was set as 100Mb for state 4 and 100kb for other states and d_i is the distance between two SNPs. The values of p are treated as unknown parameters and estimated in the Baum-Welsh algorithm. (Wang *et al.*, 2007b)

Once the model is defined, the Viterbi algorithm(Eddy, 2004) is used to infer the most likely path (state sequences for all SNPs along each chromossome).

3.6 Selection of CNV Regions

3.6.1 Quality Control

The function *filter_cnv.pl* from PennCNV returns the following quality control values: LRR standard deviation, BAF mean, BAF drift and waviness factor(WF), which are considered to filter bad samples.

As described on PennCNV tutorial, Affy data contains more noises than Illumina data, so thresholds can be set in more liberal values.

Log R Ratio is expected to have mean 0 and, for Affymetrix platforms, the standard deviation is expected to be up to 0.35 (Wang *et al.*, 2007a).

For B Allele Frequency, the mean value must be between 0.4 and 0.6. The BAF drift summarizes the deviation of BAF from the expected values of two copies (0, 0.5 and 1) (Marenne *et al.*, 2012), PennCNV suggest a threshold of 0.01, although some works with Illumina platforms set this value to 0.002.

The waviness factor identifies samples with LRR that is not consistent across the genome, the WV must be between -0.4 and 0.4 (Marenne *et al.*, 2012).

3.6.2 Minimal Regions

The output from PennCNV has one file for each sample containing different CNV regions and this information cannot be used to any association analysis. The solution to define common variables to each sample is to detect the overlap regions from all samples.

Suppose we are analyzing 6 samples (from A to F) in the region from positions 1 to 55 of chromosome 5, Table 3.2 represents this fictional example as the merged outputs from PennCNV and Figure 3.6a illustrates it. In this case, sample A has two CNVs detected and D has no CNVs.

Minimal regions will be defined using all start and end positions of all CNVs detected. For example, as shown in Figure 3.6b, "CNV 2" is defined between positions 5 and 10, because in samples C and E, the CNV starts at position 5 and, in sample A, the CNV starts at position 10.

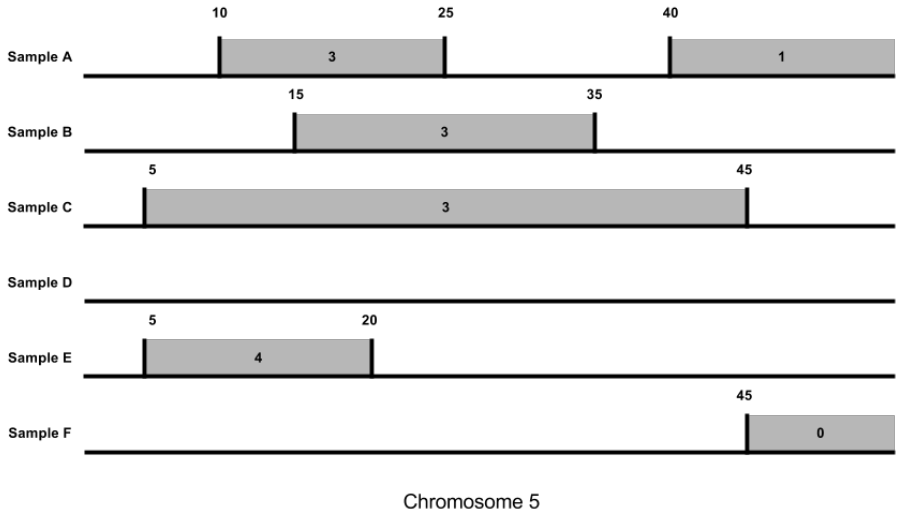
Table 3.2: *Fictional example of the merged and cleaned outputs from PennCNV. Each row represents a CNV and describes in which sample it was found, where it starts and ends, its length and its copy state (described in Table 3.1).*

ID	Chromosome	Start	End	Length	Copy State
A	5	10	25	15	5
A	5	40	55	15	2
B	5	15	35	20	5
C	5	5	45	40	5
E	5	5	20	15	6
F	5	45	55	10	1

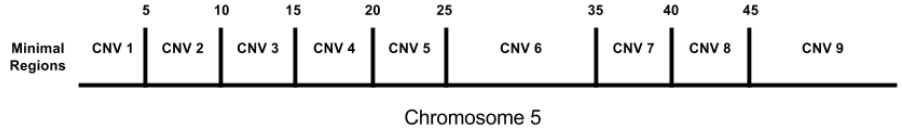
Once the minimal regions are defined, it is possible to check the copy number of these regions for each sample, as shown in Table 3.3. For this procedure, the package *CNTools* is used (Zhang, 2017) and the script of this example can be found in Appendix 4.3.5.

3.6.3 Filtering CNV Regions

After the procedure of finding the minimal regions, it is expected regions with mutations in very few samples or no mutations at all. For this reason, a procedure of excluding these regions was



(a) *Representation of CNVs from Table 3.2. The line indicates the chromosome 5 from position 1 to 55. The shadowed regions indicate the CNV regions with its respective number of copies.*



(b) *The overlap of all regions generates the minimal regions. In this case, from the 6 CNVs, we obtained 9 regions.*

Figure 3.6: *Representation of the procedure to find regions of consensus among samples.*

Table 3.3: *Copy number of each sample for all minimal regions.*

Chromosome	Start	End	A	B	C	D	E	F
5	1	5	2	2	2	2	2	2
5	6	10	2	2	3	2	4	2
5	11	15	3	2	3	2	4	2
5	16	20	3	3	3	2	4	2
5	21	25	3	3	3	2	2	2
5	26	35	2	3	3	2	2	2
5	36	40	2	2	3	2	2	2
5	41	45	1	2	3	2	2	2
5	46	55	1	2	2	2	2	0

developed.

Similar to minor allele frequency (MAF) in SNP data, which refers to the frequency at which the least common allele occurs in a given population(Consortium, 2005), it was defined a "minor CNV frequency (MCF)" of 0.02. Thus, given that a CNV region can have up to 5 groups based on number of copies (0, 1, 2, 3 and 4), the CNV will only be accounted if it have at least two groups with more than 2% of the samples.

For example, in Table 3.3, the CNV from position 1 to 5 would be disregarded once it has only one group (all samples has 2 copies, then group 2 has 100 of the samples). The script used for this step can be found in Appendix 4.3.5.

3.7 Heritability Estimation and Phenotype Association

The starting point for understanding the statistical methodology of heritability estimation is the linear mixed model, which is described in Equation 3.10, where $\mathbf{Y}$ is the vector $(n \times 1)$ containing the response variable with $n = \sum_{i=1}^{c} n_i$ and c the number of clusters, β is vector $(p \times 1)$ with the fixed effects parameters, X and Z are, respectively, $(n \times p)$ and $(n \times q)$ known model specification matrices of full rank, γ is a $(q \times 1)$ random effects vector, and ϵ is a $(n \times 1)$ vector of random (within-units) errors(Duarte *et al.*, 2014).

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon \tag{3.10}$$

Assuming γ and ϵ are uncorrelated with $E[\gamma] = 0$, $E[\epsilon] = 0$, $Cov[\gamma] = \sigma^2\mathbf{D}$ and $Cov[\epsilon] = \sigma^2\mathbf{R}$, where $\mathbf{R}$ and $\mathbf{D}$ are positive definite matrices with known structure, the Equation 3.11 holds.

$$Cov[Y] = (ZDZ^T + R)\sigma^2 \tag{3.11}$$

For all quantitative phenotypes, the variance of a given trait in the population can be explained by biological and environmental factors, as well as their interactions. Based on this, the heritability is the proportion of a phenotypic variance that is due to the genetic factor(Visscher *et al.*, 2008). Hence for the heritability estimation, the Equation 3.10 can be rewritten as the variance component model (Equation 3.12), a common tool for family-based association analysis (Almasy e Blangero, 1998):

$$y_i = \mu + \beta_j x_{ij} + g_i + \epsilon_i, \tag{3.12}$$

where, y_i = phenotype of $i-$th individual, μ = overall mean factor, β = regression coefficients associated to the covariates, g_i and ϵ_i are the polygenic random effect and the error component, respectively, assumed to be uncorrelated and normally distributed with mean zero and variance σ_g^2 and σ_e^2, respectively, such that $\sigma_g^2 + \sigma_e^2 = \sigma^2$ is the phenotypic variance.

For family data, the covariance matrix is $\Omega_f = 2\phi_f\sigma_g^2 + I_f\sigma_e^2$, in which, $2\phi_f$ is the kinship matrix, containing the coefficients of relationship between the individuals i and i' from family f, given by $\left(\frac{1}{2}\right)^r$ where r represents the degree of relationship and I_f the identity matrix. The heritability is defined as intraclass correlation coefficient, given by:

$$H^2 = \frac{\sigma_g^2}{\sigma^2} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \tag{3.13}$$

The null hypothesis of interest in this case can be described as Eq.3.14.

$$H_0 : \sigma_g^2 = 0, \tag{3.14}$$

which can be tested by using Likelihood Ratio Test (LRT) (Self e Liang, 1987). This method tests whether the additive polygenic effect in each analysis accounted for a significant component of the