Variation in copy number on the genome of the Brazilian population

Ana C. M. Ciconelle[1], Júlia M. P. Soler[1], Alexandre C. Pereira[2]
[1]Institute of Mathematics and Statistics - University of Sao Paulo – USP, Brazil
[2] Heart Institute - University of Sao Paulo – USP, Brazil

Abstract

Copy number variation (CNV) is an alteration in the number of copies of a DNA segment, unbalancing the diploid state in humans at any given locus on the genome. The CNV region can include from a single nucleotide polymorphism (SNP) to several genes, and such variation can be classified in five states: 0 (deletion of two copies), 1 (deletion of one copy), 2 (normal state), 3 (single copy duplication) and 4 (double copies duplication). Several diseases (such as uric acid, pancreatitis and nervous system disorders) and phenotypes (such as height and cholesterol levels) have been associated to this kind of structural variation, suggesting that inheritance patterns can be involved besides revealing variability across populations. In this study we propose a pipeline for CNVs calling from SNP array data. Further, in collaboration with Heart Institute (USP), this work uses dataset from Baependi Heart Study to characterized the CNVs in the Brazilian population and associate them with height. Genomic and phenotype data consisted of 1,120 related individuals sampled according to family-based design. The results pointed out to CNV regions specific for Brazilian population, but also for similarities with others populations according the length and number of CNVs in samples. In addition, based on trios data (parents and offspring) it was observed that the CNV transmission could not follow the Mendelian laws. Our work also identified a region in the chromosome 9 associated to height, where it carries a duplication with an expected height dropped by approximately 3cm.

Keywords

CNV calling; association studies; height, missing heritability; mixed model

1. Introduction

As described by Lewis (2012), Genome Wide Association Studies (GWAS) aim to associate genetic markers, candidate genes or genome regions with complex traits and diseases, which are likely derived from multiple genes and environment, such as height and diabetes. In addition, discovering the associations between diseases and genetic factors is an important step to understand the pathogenesis of the diseases and to facilitate the process of diagnosis and treatment. The most used genetic variant for GWAS is the single nucleotide polymorphism (SNP), but other variants, as small

insertions/deletions (indels) and copy-number variations (CNVs), are also available.

Several studies are being performed to catalogue the human genetic variants to facilitate GWAS, such as the HapMap Project and 1000 Genomes Project. In both projects, samples are majorly from African, Asian and European populations and they aim to identify genetic variants with frequencies of at least 1% in the studied populations, including not only SNPs, but also structural variants and small insertions/deletions (The International HapMap Project, 2003; 1000 Genomes Project Consortium, 2016). Even though there is a major success in gene discovery, the percentage of variance explained by GWAS loci for many traits is relatively low. Thus, a substantial part of the traits variation is still unexplained. This phenomenon is called missing heritability. One example of trait with a high missing heritability is the height. In Manolio et al. (2010), two of the solutions cited to revealing the missing heritability is to use different types of genetic variants including common and rare variants.

Based on these scenarios, in this work, our focus is on CNV detection since this kind of variant is not as well characterized as SNPs, but it is expected to have an important role on the association with several traits and diseases. Copy number variation occurs when the number of copies of a particular region (one or more loci) of the DNA differs from two in autosomes or one/two in allosomes and can to explain phenotypic variability in humans. The effects of CNVs to human diseases are not yet well known although several diseases have been associated to this kind of polimorphism, such as uric acid (Scharpf et al., 2014).

GWAS are usually based on reference maps which do not take into account the population-specific and rare variants. In addition, Sanna et al. (2011) shows that adding rare variants in association studies doubled the explained heritability of traits. Therefore, identifying different types of variants and including data from specific populations can explain the missing heritability of traits and diseases. This motivates to build genomic reference maps for specific populations, as proposed by the project Genome of Netherlands (Boomsma et al., 2014), which aims to characterize genetic variants from Dutch population, including rare variants.

Considering the unknown influence of CNVs on anthropometric measurements and the lack of studies based on Brazilian population, this work was developed in collaboration with the Laboratory of Genetics and Molecular Cardiology (Heart Institute-USP, Brazil). Using the database from the Bapendi Heart Study described by Egan et al. (2016), we analysed the genotype (SNP data) and phenotype data from 80 families to characterize the CNVs in the Brazilian population and to understand their association with phenotypes, such as height. The main purpose of this project is to present methodologies

to quantify and call CNVs from SNP platforms and to analyse such data considering family based designs and characterize the patterns of the CNVs detected in this population.

2. Methodology

Dataset

Due to multiple waves of immigration, Brazil has a highly admixed population, which can be driven by genetic and environmental influences on several traits. The Baependi Heart Study is being conducted by the Heart Institute since 2005 to develop a longitudinal family-based cohort study for understanding the variation of cardiovascular risk factors within the Brazilian population and disentangle its genetic and environmental components. The data provides information about 105 families (1,666 individuals, 723 male and 943 females) living in the village of Baependi, in the state of Minas Gerais, Brazil. Data from 631 nuclear families were available, with offspring ranging from 1 to 14. 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.

For each participant a questionnaire was used to obtain information regarding family relationships, demographic characteristics, medical history and environmental risk factors. Anthropometric measures, physical and clinical examination and electrocardiogram of the participants were performed by trained medical students. Genomic DNA was extracted by standard procedures. From DNA samples, genotyping with SNP array was made based on Affymetrix Platform 6.0 and 1,120 CEL files were obtained, which stores the intensity values of each probe array for a single sample and several others information. More details are described in Egan et al. (2016).

Overview

The methodology used in this work is summarized by Figure 1, which describes the pre-processing of SNP data, the CNV calling and the CNV analysis. For the pre-processing of SNP data and the CNV calling, the software Affymetrix Power Tools (APT) (Affymetrix, 2017), PennCNV by Wang et al. (2007) and packages from the R environment were used. Using APT, given the CEL files, signal intensity values for probes are normalized through quantile normalization. Then, the median polish is applied to get the final cleaned intensity values for alleles A and B for each SNP. Also, the individual genotype calls is made using the Birdseed algorithm. For each SNP in each sample, the genotype will be coded as 0, 1 and 2 for AA, AB, BB and -1 for missing values, respectively, with its corresponding confidence scores. In addition, a final report will infer the sample sex. PennCNV generates canonical genotype clustering files based on the output files from APT. These files contain cluster

positions of each SNP for each canonical genotype (AA, AB and BB). Then, the calculation of LRR and BAF values for each SNP and each sample are made. All these values are used by a hidden Markov model (HMM) to CNV calling for each sample. Quality control values are also generated. The identified CNV regions are specific for each sample (individual). We excluded the samples that do not pass in the quality control. Then, a new set of minimal regions, defined by the overlap regions across all samples, was built and all minimal regions with a low frequency of CNVs (less than 2%) were removed. The final minimal regions are then ready for the CNV analysis of this work.

The SOLAR package combined with R scripts was used to analyse the CNVs, calculate the heritability of traits and associate CNVs and traits. Heritability corresponds to the intraclass correlation coefficient defined under linear mixed model formulation and considering family-based designs.

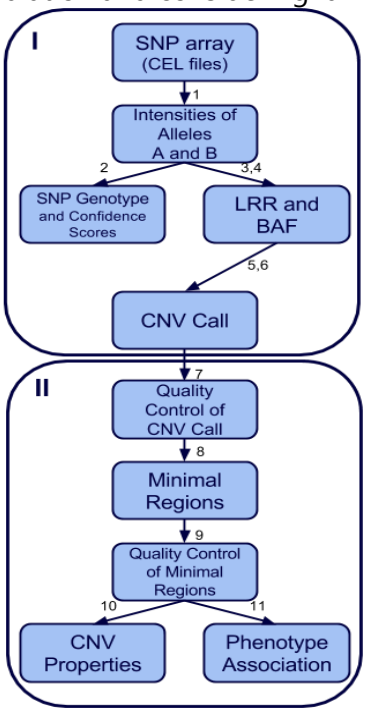


Figure 1: Flowchart of the pipeline. Box I indicates CNV calling and box II indicates CNV analysis. The number indicates which function was used: 1 (apt-probeset-summarize) and 2 (apt-probeset-genotype) are from APT, 3 (generate_affy_geno_cluster.pl), 4 (normalize_affy_geno_cluster.pl), 5 (kcolumn.pl), 6 (detect_cnv.pl) e 7 (filter_cnv.pl) are from PennCNV, 8 (CNTools) is from bioconductor package cntools, 9, 10 and 11 are basic functions from R environment and SOLAR software.

3. Results

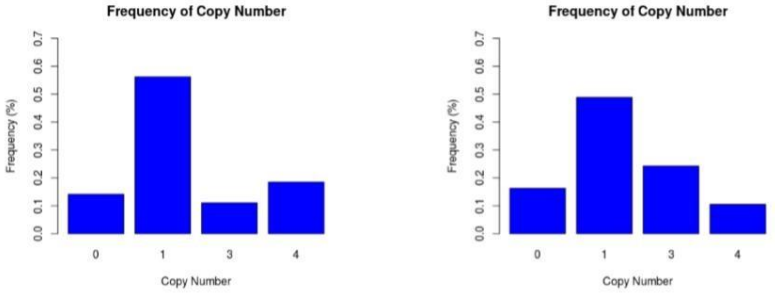
By the end of the CNV calling, each sample has a file describing the identified CNVs as showed in Table 1. From the 1,120 samples 910 were

considered for analysis due the quality control filtering. From the original data, we were able to identify 375,312 CNVs and, after the cleaning procedure, this value dropped to 135,414 CNVs. From these CNVs, we obtained 64,107 minimal regions, in which we considered the overlap of CNVs. Due the low frequency of some CNVs in the samples, after filtering, only 8,794 were considered.

How many CNVs does an individual have? For this question, the number of CNVs we obtained from each sample varies from 17 to 2,921 CNVs. However, we also can observe that a subgroup of 83% of the samples contains less than 100 CNVs, which is expected limit for PennCNV. For this subgroup of samples, the mean number of CNVs per sample is 56.49 (standard deviation equal to 15). For both, the complete samples and the subgroup, the median of 60 and 57 CNVs, respectively, are compatible with similar studies. We also can observe that deletions are more frequent than duplication as show in Figure 2.

Sample	Chr	Start	End	Number	Length	State	CN	First Marker	Last Marker
1	15	22231485	22264715	31	33231	2	1	CN_691574	CN_691602
1	19	59989695	60040503	29	50809	2	1	CN_170378	SNP_A-4271224
1	17	18296117	18373803	21	77687	2	1	CN_749706	CN_751779
1	17	67057139	67076931	22	19793	2	1	CN_744214	CN_744222
1	9	44181813	44569219	23	387407	2	1	CN_1322576	CN_1322482
1	4	64380064	64390853	30	10790	1	0	CN_1052052	CN_1052079

Table 1. Illustration of the file containing the CNVs from sample 1. PennCNV generates a file with this structure for each sample. Each line describes a CNV. Columns "Chr", "Start" and "End" indicates the region of the CNV. "Number" is the number of markers from the Affymetrix 6.0 platform inside the region of the CNV. "Lenght" is the size of CNV in base pairs (bp). "State" corresponds to HMM states and "CN" is the number of copies associated to the state. "First and Last Markers" identify the markers where the CNV starts and ends.



(a) *All samples* (b) *Samples with less than 100 CNVs*

Figure 2. Distribution of CNVs regarding the number of copies. 0 and 1 indicate deletion, while 3 and 4 indicate duplication. (a) contains CNVs from all the samples and (b) contains only CNVs from samples with less than 100 CNVs.

How long are the CNVs identified in the Brazilian population data? According our results, the length of a CNV varies between 3bp to 27,435,314bp (27.5Mb) and follows a log-normal distribution as obtained by Scharpf et al.

(2014). Figure 3 shows histograms of the size of CNVs, indicating that deletions are, in general, shorter than duplications.

Where are the CNVs? The literature shows that chromosomes 19, 22 and Y present the biggest proportions of CNVs. From our dataset, chromosomes 19 and 8 have more regions of CNVs based on the number of base pairs. However, when only CNVs detected in at least 5% of the samples are considered, chromosomes 19 and 9 have the biggest proportions.

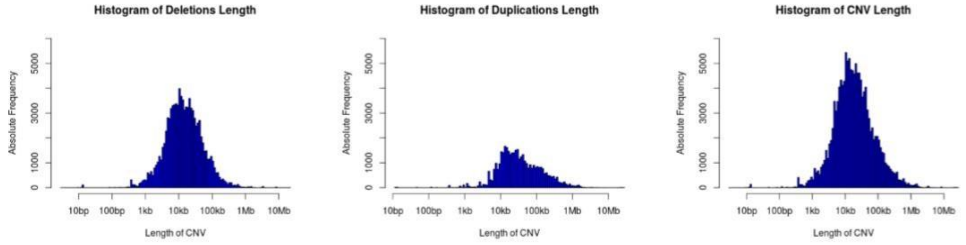


Figure 3. Histograms of CNV length. Figure (a) considers only deletions; Figure (b) considers only duplications and (c) contains all CNVs. The data is presented in exponential scale.

For understanding the distribution of CNVs along the genome, Figure 4 shows the absolute frequency of the detected CNVs along the positions on the chromosome 1 and 6. The region with highest presence of CNVs across samples is in chromosome 1, between positions 72,541,505bp and 72,583,736bp (Figure 4), which, on average, 818 samples from the total of 910 has a deletion or duplication.

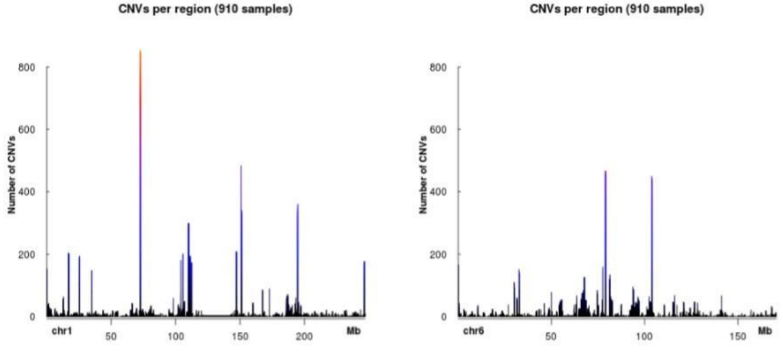


Figure 4. Number of CNVs per region after finding the minimal regions. The x-axis represents the positions of the chromosome by base pairs. The y-axis indicates the number of CNVs detected in the respective position for 910 samples.

Does the CNV follow Mendelian laws? This question is related to the inheritance pattern of the CNVs. At total, 106 trios were analysed from the Baependi families. They include trios with the same parents with different offsprings. As expected, normal parents and normal offspring is the most

common combination of CNV occurrences, in which, on average, 77.45% of the trios are all normal for all 8,794 CNVs. When we consider the case of one parent being normal and another having a deletion, we expected, under the Mendelian law, that the proportions of children with two copies would be similar to the children with one deletion. However, on average, 7.52% of the trios has parents with one normal parent and another with single deletion and the mean frequency of the trios with offspring with deletion is 1.29%, while with normal offspring is 6.06%. It means that, in general, the affected parent transmits preferentially the normal allele instead of the allele with deletion. A similar situation can be found for trios in which one of the parents is normal and the other has one duplication.

Are the CNVs associated to height? Height is a complex phenotype and its heritability is estimated to be around 80%. Due to the missing heritability, we explored this phenotype in association with CNVs. A linear mixed model was applied considering height as response variable and age, sex, ancestry coefficients and CNV as covariates. We adjusted the model in three different ways, in which the CNV was considered as dichotomous, having linear effect and as categorical covariate in five levels. Based on the results, the region from 78,960,219bp to 78,967,224bp in chromosome 9 is the most significant among the CNVs. Figure 5 shows the Manhattan plot for the model with CNV as a discrete variable. Also, the results indicate that the presence of duplication decreases the expected height by approximately 3cm as show in Figure 6.

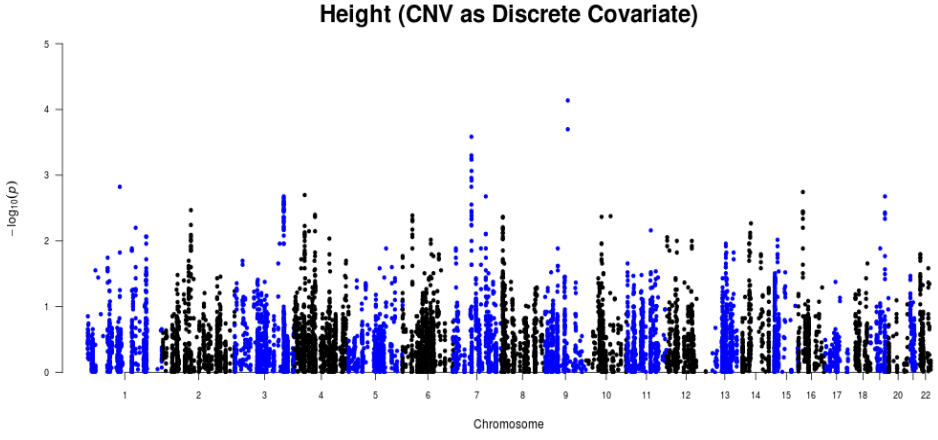
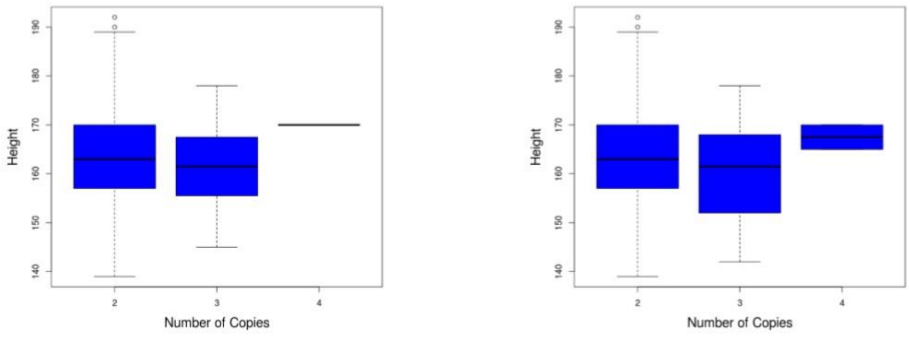


Figure 5. Manhattan plot from the second model. y-axis indicates the –log10 (p-value) of each CNV in association with height. The position used in the x-axis is the center position (in bp) of the CNV.



(a) (b)

Figure 6. Distribution of the height. These distributions are based on the number of copies for the regions from 78,960,219bp to 78,961,487bp **(a)** and from 78,961,488bp to 78,967,224bp **(b)** of chromosome 9.

4. Discussion and Conclusion

Our approach allows us to characterize CNVs occurring on Brazilian population. The CNV database built in this work can be used for association studies with different phenotypes. From this descriptive analysis of the obtained CNVs, we could observe that the distribution of the length and the number of CNVs per sample are similar to other populations as described in the literature, but specific CNV regions were also identified. The minimal regions identified can be considered as genetic markers specific for the Brazilian population. Further work can be performed based on the CNV database obtained and the annotation of the common identified CNVs should be made for better understanding the biological system.

References

1. 1000 Genomes Project Consortium (2016). HHS Public Access. *Nature* 526(**7571**): 68–74.
2. Affymetrix (2017). Affymetrix Power Tools: MANUAL: apt-probesetsummarize (1.20.0).
3. Boomsma et al. (2014). The Genome of the Netherlands: Design, and project goals. *European Journal of Human Genetics* 22(**2**):221–227.
4. Egan et al. (2016). Cohort profile: the Baependi Heart Study—a family-based, highly admixed cohort study in a rural Brazilian town. *BMJ Open* 6(**10**):e011598.
5. Lewis C. M., Knight J. (2012). Introduction to genetic association studies. *Cold Spring Harbor Protocols* 2012(**3**):297–306.
6. Manolio T. A. et al. (2010). Finding the missing heritability of complex diseases. *Nature* 461(**7265**):747–753.
7. The International HapMap Consortium (2003). The International HapMap Project. *Nature* **426**:789– 796.
8. Sanna et al. (2011). Fine Mapping of Five Loci Associated with Low-Density Lipoprotein Cholesterol Detects Variants That Double the Explained Heritability. *PLoS Genetics* 7(**7**).
9. Scharpf R. B. et al. (2014). Copy number polymorphisms near SLC2A9 are associated with serum uric acid concentrations. *BMC Genetics* 15(**1**):1–13.
10. Wang et al.(2007). PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* 17(**11**):1665–1674.