

MSc in Bioinformatics

Master Thesis

**Detection of obesity susceptibility genomic variants in Spanish population using sequencing data**

Isaac David De la Hoz Saltaren

*Supervised by*

Juan R González, ISGlobal

*Academic tutor*

Raquel Egea, UAB



July 2019

# Acknowledgments

# Approval and signature

# Abstract

# Table of contents

[Acknowledgments 3](#_Toc11339068)

[Approval and signature 4](#_Toc11339069)

[Abstract 5](#_Toc11339070)

[Table of contents 6](#_Toc11339071)

[1. Introduction 8](#_Toc11339072)

[1.1. Obesity 8](#_Toc11339073)

[1.1.1. Obesity in Spanish population 8](#_Toc11339074)

[1.2. Body weight control and causes of obesity 8](#_Toc11339075)

[1.2.1. Body weight control 8](#_Toc11339076)

[1.2.2. Pathway of energy homeostasis 9](#_Toc11339077)

[1.2.3. Disorder of energy homeostasis 10](#_Toc11339078)

[1.3. Importance of genetics in obesity 10](#_Toc11339079)

[1.4. Obesity susceptibility variants 10](#_Toc11339080)

[2. Objectives 12](#_Toc11339081)

[3. Material and Methods 13](#_Toc11339082)

[3.1. Data description 13](#_Toc11339083)

[3.1.1. Samples 13](#_Toc11339084)

[3.1.2. Methodology of DNA extraction and sequencing 14](#_Toc11339085)

[3.1.3. Genomic alignment 14](#_Toc11339086)

[3.2. SNV detection 14](#_Toc11339087)

[3.2.1. Variant Calling and annotation (GATK) 14](#_Toc11339088)

[3.2.2. Statistical analysis 16](#_Toc11339089)

[3.3. CNV Analysis 19](#_Toc11339090)

[3.3.1. CNV prediction. 19](#_Toc11339091)

[3.3.2. Statistical analysis 21](#_Toc11339092)

[4. Results 23](#_Toc11339093)

[4.1. SNV detection. 23](#_Toc11339094)

[4.1.1. Potential causative SNV 23](#_Toc11339095)

[4.1.2. Enrichment analysis. 24](#_Toc11339096)

[4.1.3. Variant selection and overlapping with existing databases. 25](#_Toc11339097)

[4.2. CNV analysis 27](#_Toc11339098)

[4.2.1. Counts of reads. 27](#_Toc11339099)

[4.2.2. ExomeCopy model 27](#_Toc11339100)

[4.2.3. Summarizing individual CNVs (CNV Ranger). 28](#_Toc11339101)

[4.2.4. Association analysis. 28](#_Toc11339102)

[5. Discussion 29](#_Toc11339103)

[6. Conclusion 30](#_Toc11339104)

[7. Bibliography 31](#_Toc11339105)

# Introduction

## Obesity

Obesity is a medical condition that is defined as excessive accumulation of fat that is sufficient to adversely affect health1,2. According World Health Organization (WHO), people with a body mass index (BMI; weight in kg/height in m2) higher than 30 kg/m2 are considered obese and higher than 25 kg/m2 are considered overweight. The 30% of Americans and 10%–20% of Europeans are classified as obese, with the prevalence rising in many developing countries1. Being overweight or obese can have a serious impact on health. Carrying extra fat leads to serious health consequences such as cardiovascular disease (mainly heart disease and stroke), type 2 diabetes, musculoskeletal disorders like osteoarthritis, and some cancers (endometrial, breast and colon). These conditions cause premature death and substantial disability3.

### Obesity in Spanish population

The 26.6% of adult population in Spain are considered as obese and 62% are overweight. Furthermore, in Spanish childhood population, the prevalence of obesity has been increasing in recent years to such a stent that, the prevalence has increase until 18.3% among children (0-9 years of age) and 30% among adolescents (10-19 years of age). This data has made Spain be considered by WHO as one of the countries with the highest prevalence of obesity and overweight4.

## Body weight control and causes of obesity

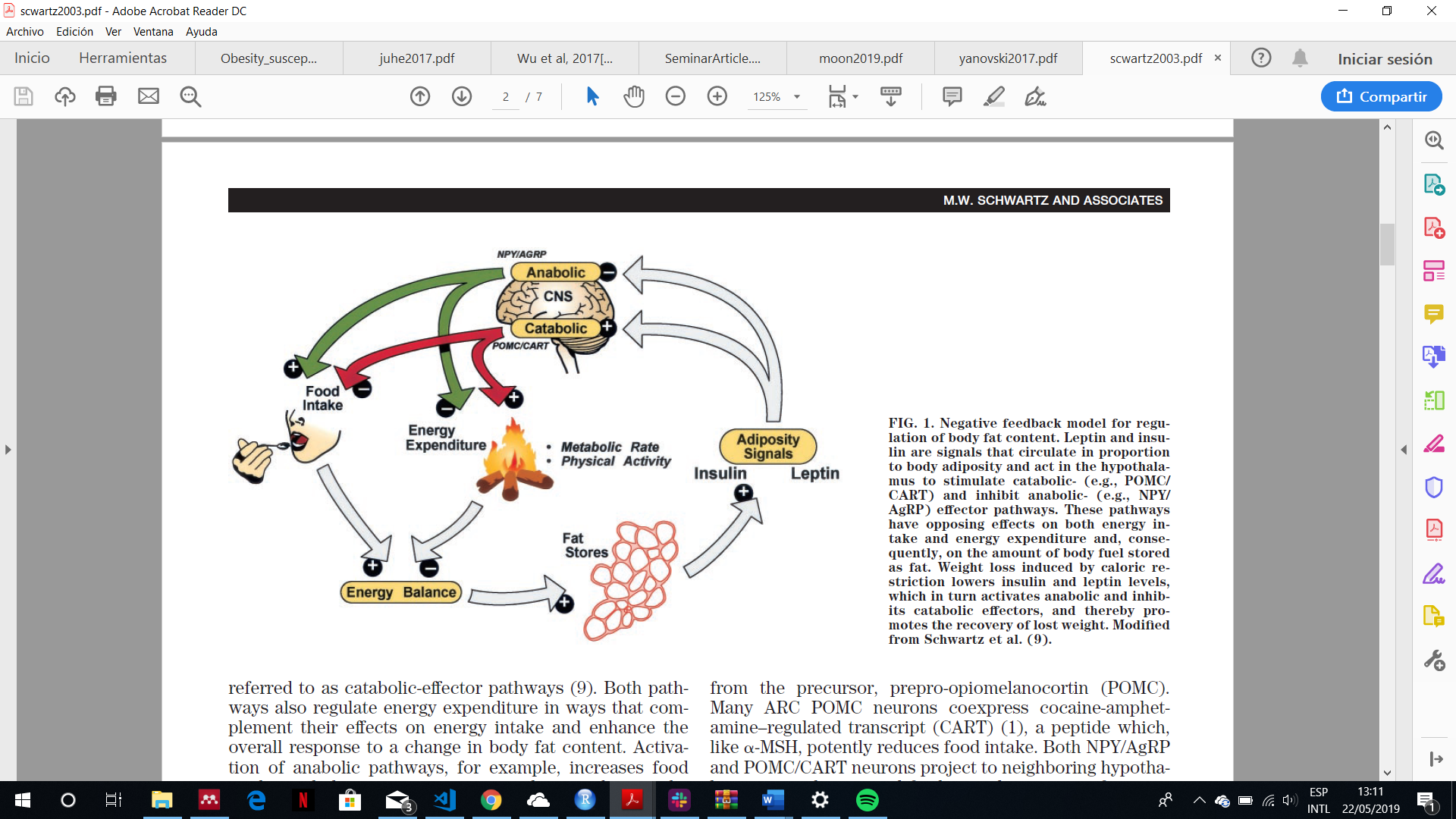
### Body weight control

The humans are able to regulate their body weight over long periods of time despite day-to-day variation in the number of calories consumed and in levels of energy expenditure1. The maintenance of constant body weight and body composition requires two conditions be met. [1] An even energy balance must be attained, i.e. energy expenditure must on average be equal to energy intake. [2] There must be an even balance for each individual substrate, i.e. protein, carbohydrate and fat oxidation must be equal to protein, carbohydrate and fat intakes respectively. If this state were not present, body composition would inevitably change, even during isoenergetic feeding. If energy intake continuously exceeds energy expenditure, the excess energy ingested has to be deposited in order to increase the nutrients stores5.

There is clearly a ‘hierarchy’ in substrate oxidation during overfeeding. Any increase in protein intake will rapidly lead to stimulation of protein oxidation, restoring a steady protein balance. The same is true for carbohydrates, the oxidation of which increases over 1–3 times to match any increase in carbohydrate intake. The result of this hierarchy is that excess energy intake leads essentially to fat storage, mainly in subcutaneous and visceral adipose tissue. In contrast, a period of hypoenergetic feeding will lead to a negative fat balance and a loss of adipose tissue1,5.

### Pathway of energy homeostasis

Pathways that stimulate food intake and promote weight gain are referred to here as anabolic-effector pathways, whereas those that promote anorexia and depletion of body fat are referred to as catabolic-effector pathways6. Both pathways also regulate energy expenditure in ways that complement their effects on energy intake and enhance the overall response to a change in body fat content. Activation of anabolic pathways, for example, increases food intake and decreases energy expenditure, whereas the reverse is true for catabolic pathways6. Anabolic and catabolic pathways are generally regulated in a reciprocal manner, such that increases in the activity of one are often accompanied by decreases in the other6. The anabolic and catabolic pathways sense changes in energy balance due to the hormones leptin and insulin that circulate on blood proportionate to body fat mass and enter into the brain, where they bind to and activate their respective receptors on the plasma membrane of targets neurons6. Low concentrations of leptin and insulin increase energy intake and reduce energy expenditure. Hence, the reciprocal nature of the neuronal response to an energy deficit (activation of anabolic pathways and inhibition of catabolic pathways) may be accounted for, at least in part, by reduced levels of these two hormones (Figure 1)1,6,7.



**Figure 1: energy balance pathway.** This model explains the body fat mass storage mechanism. When the food intake increases, the excess of energy is stored at adipocytes. Once the adipocytes have stored energy in form of fat, they produce leptin that activate central nervous system pathways which it stimulates the decrease of energy intake (satiety signals) and the increase in energy expenditure. Figure adapted from Schwartz et.al 20036.

### Disorder of energy homeostasis

Disorders of energy homeostasis are fundamentally due to factors that disrupt the balance between energy intake and expenditure over time, the utilization of substrates (fat, protein, carbohydrate), and/or nutrient partitioning (storage of excess calories). The environmental influences of weight gain such as the adoption of sedentary lifestyles due to reduced physical activity at work and in leisure time coupled with an abundance of easily available, energy-rich, highly palatable foods represents a nutrition transition that, according with the WHO, is now one of the greatest factors to ill health worldwide1.

Other factor that disrupt the energy homeostasis is the genetic. In any environment, either energy rich or energy lacking, there are considerable distribution of different body weight among people. This evidence says that not only environmental factors cause obesity but also genetics factors1,2,7–9.

## Importance of genetics in obesity

The genetic contribution to body weight has been established through family studies, investigations of parent-offspring relationships, and the study of twins and adopted children10,11. These studies estimate a heritability (Fraction of the total phenotypic variance of a quantitative trait attributable to genes in a specified environment) of 40-70%. This genetic predisposition has been widely recognized in the human evolutionary history. Obesity stem from natural selection on our ancient ancestors favouring “thrifty genes”, defined as conferring a phenotype of being extremely efficient keeping all extra energy during periods of food abundance in order to deal with large famine periods. In modern society, however, with plentiful and continuous food, this thrifty phenotype process deleterious because it promotes efficient storage of fat for a famine period that never comes1,12.

## Obesity susceptibility variants

As explained before, genetics factors have big genetics influences in the appearance of the obesity. These genetic influences are likely to operate across the weight spectrum but may be more penetrant when studying childhood-onset obesity and at both extremes of the BMI distribution (thinness and severe obesity). Genetic variance of obesity depends on 5 factors1:

* The nature and amount of mutational variance in a population
* The segregation and frequency of the alleles that influence a trait in a particular population. The lower minor allele frequency the worse phenotype.
* The effect size of the variant. The effect that a variant can produce could be additive or non-additive.
* The mode of gene action
* The degree of genetic control of phenotypic variance of the trait in question.

Until now, 97 genetic loci have been discovered associated with BMI through Genome-wide association studies (GWAS) approaches13. Nonetheless, these loci only explain 2.7% of the variances of BMI. Several monogenic drivers of isolated early-onset obesity have been identified, emphasizing the importance of energy homeostasis (LEP, LEPR, POMC, MC4R) and cilia function (CEP19)8,14. In addition, the gene that encodes the fat mass associated protein (FTO) has unequivocally been associated with obesity by the existence of single nucleotide polymorphism (SNPs) in both childhood and adult obesity populations.

In the other hand, several copy number variants (CNV) that contribute with the obesity heritability have been reported. including deletions upstream of the NEGR1 gene15, proximal and distal deletions at 16p11.216,17, gains at 10q26.6 containing the CYP2E1 gene (MIM 124040)18, and homozygous deletions at 11q11 encompassing olfactory receptor genes19, among others.

In short, both single nucleotides variants (SNPs and Indels) and copy number variants may be focus of study not only for their everyday growing causal variants discovery rate, but also because the variants already found only explain less than 3% of the heritability of BMI.

# Objectives

In order to contribute to the variant discovery, this master thesis has the following objectives:

* To obtain the single nucleotides variants (such as SNPs or INDELs) and copy number variants from exome sequencing data through developing appropriates workflows with that aim.
* To develop appropriate statistical analyses in order to discover new single nucleotide variants correlated with obesity in Spanish population.
* To infer CNVs regions from the CNVs calls found in exons.
* To find the biological significance of these variants in developing obesity analysing the genes and metabolic pathways affected.

The data used for performing the whole analyses come from Spanish individuals with extreme obesity. By this way, the power for finding new significant variants is upper.

# Material and Methods

## Data description

### Samples

The data come from 18 unrelated adult individuals with severe early-onset obesity of Iberian origin (their characteristics are exposed in **Table 1**). Of these, 10 individuals were recruited from the Endocrinology Service of the General Hospital of Valencia (Spain), 4 adult individuals from the Endocrinology Service of the Infanta Cristina Hospital in Badajoz (Spain), and 4 adult individuals from the Coimbra Hospital (Portugal) between May and December 2017. The selection criteria were the following:

* More than 3 Kg at be born
* To get the obesity being less than 6 years old
* BMI higher than 45 kg/m2 in adults and greater than the 99th percentile in children
* The existence, at least, other three cases of morbid obesity among first- or second-degree relatives
* To be free of hypertension, diabetes or any cardiovascular disease

**Table 1:** Anthropometrics characteristics of the sample

|  |  |  |  |
| --- | --- | --- | --- |
|  | Man | Woman | total |
| Sex | 9 | 9 | 18 |
| Years-old | 45.2 ±4.1 | 35.2 ±10.7 | 40.2 ±9.3 |
| Weigh(kg) | 175.7 ±31.6 | 130.2 ±19.9 | 153 ±34.6 |
| Height (m) | 1.76 ±0.07 | 1.63 ±0.03 | 1.69 ±0.08 |
| BMI (kg/m2) | 56.96 ±11.3 | 49.4 ±8.5 | 53.2 ±10.2 |

***Important considerations:***

* *The study protocol was approved by the Directorate General of Innovation and Curriculum Development and the Ethics Committee of the Ministry of Education (both of the Government of Portugal), and by the Ethics Committee of the General Hospital of Valencia and the Hospital Infanta Cristina de Badajoz (Spain).*
* *The study was conducted in accordance with the institutional and ethical requirements of the University of Coimbra, as well as the Declaration of Helsinki and its subsequent revisions. The written informed consent of all the patients was obtained before the participation in the study.*

### Methodology of DNA extraction and sequencing

The genomic DNA was extracted from peripheral blood mononuclear cells using the MagNA pure system (Roche Life Science, Barcelona, España), according with the manufacturer’s instructions. The quantification and purity of DNA was determined by the fluorometer Qubit 2.0 (Thermo Fisher Scientific Inc., Waltham, MA, EE. UU.) and the spectrophotometer NanoDrop (Thermo Fisher Scientific Inc., Waltham, MA, EE. UU.) respectively. The DNA integrity was observed using agarose gel electrophoresis.

From each sample, genomic DNA was broken into 150-200 base pair (bp) fragments using the focused-ultrasonicator Covaris S2 (Covaris, Brighton, Reino Unido). The exome capture was prepared using the instructions provided by the Agilent SureSelect Human All Exon V6 (Agilent Technologies, Santa Clara, CA). The sequencing was performed by the platform Illumina HiSeq 2500 (Illumina, Inc., San Diego, CA, EE. UU.) using v3.0 SBS with densities of flow grouping per cell of 700-800 K/mm2 on average.

### Genomic alignment

Before genomic alignment, the reads were pre-processed in order to eliminate the adapters used for performing the sequencing step as well as low quality reads through the software *FastQC* version 0.10.1 and *Cutadapt* version 1.8.1. Next, the reads were aligned in front of the genome version GRCh38/hg38 through the software BWA (version 0.7.12) and the duplicates were eliminated through Picard (version 1.92). Finally, an alignment quality control was performed through the software *Qualimap* version 2.1.

## SNV detection

From BAM files, a workflow was created using different tools in order to, firstly, to obtain all SNV (SNPs and INDELs) and, finally, to perform statistical analyses in order to find variants significantly associated with obesity (Figure 2). All parts of the workflow are explained in the following sections. (The complete code is available in the following GitHub repository https://bit.ly/2ELJ6R4).

### Variant Calling and annotation (GATK)

The variant calling procedure was performed using GATK20 tools following the best practices protocol recommended by GATK21,22. This procedure was divided in 5 parts that are explained as follow.

* ***Haplotype caller*.** The tool used for call the variants from the obese samples was *Haplotype Caller* in *GVCF* mode*.* This tool is capable of calling SNPs and indels simultaneously via local de-novo assembly of haplotypes in an active regions21.

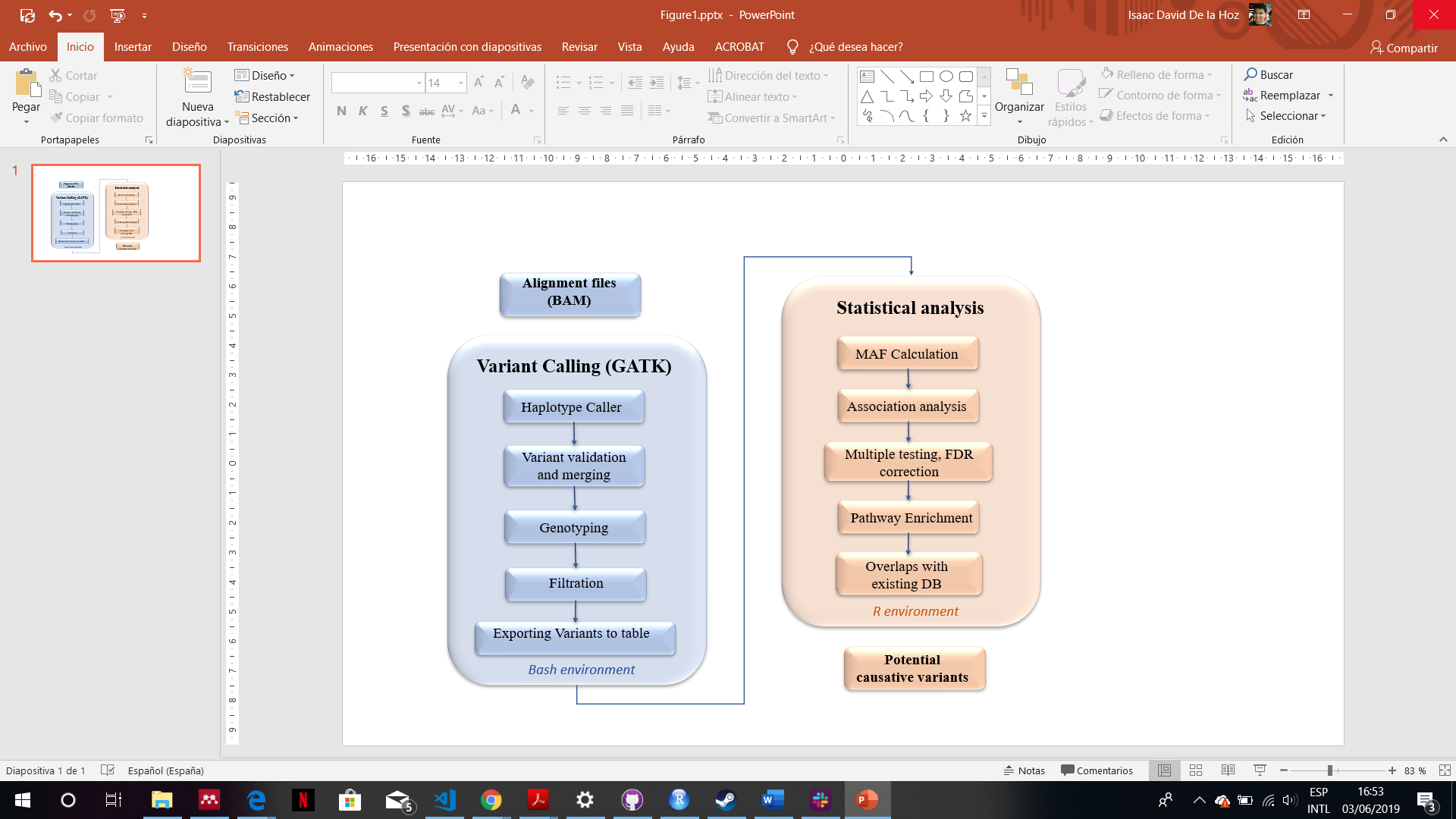
For the later statistical analysis was needed to calculate the minor allele frequency of each allele, and, for doing that, all the variants called from all samples had to be allocated in a single multi-VCF file. Therefore, according the best practices protocol21, the *haplotype caller* was run to every single sample activating the **GVCF mode** for generating genomic VCF files (gVCF) that can be merged. This mode allowed the program to be able to produce VCF files containing the information about every position in the genome regardless of whether a variant was detected at this site or not (gVCFs). In addition, additional information such as genotypes likelihoods and genotype quality are generated to improve the next merging and genotyping steps.

* **Variant validation and merging**. Before merge, the gVCF files obtained from the samples were validated using the GATK tool *ValidateVariants.* This step was performed just in case any gVCF file had any format error because if there were any, it would fail the merging step. Once validated, all gVCF were merged through the tool *CombineGVCFs* obtaining as a result a multi-sample VCF file.
* **Genotyping.** One the multi-sample VCF was created, it was needed to perform joint genotyping in order to assign the alleles of every single record, taking into account the genotype likelihoods generated by *HaplotypeCaller*. The tool used was *GenotypeGVCFs.*
* **Filtration and exporting to table.** The philosophy of the GATK is to produce a large, highly sensitive callset. This make some low-quality variants to be in the VCF file. For this reason, the output needed to be refined but through additional filtering21. For this reason, the VCF file was filtered applying the recommended thresholds21 exposed in Table 2.

**Table 2**: filtering parameters applied to VCF file21. QD: Variant call confidence normalized by depth of sample reads supporting a variant (QualByDepth), MQ: Root Mean Square of the mapping quality of reads across all samples (RMS Mapping Quality), FS: Strand bias estimated using Fisher's Exact Test, SOR: Strand bias estimated by the Symmetric Odds Ratio test, MQRankSum: Rank Sum Test for mapping qualities of REF versus ALT reads, ReadPosRankSum: Rank Sum Test for relative positioning of REF versus ALT alleles within reads, InbreedingCoeff: Likelihood-based test for the inbreeding among samples.

|  |  |  |
| --- | --- | --- |
| Parameters | Fr SNPs | For Indels |
| QD | < 2.0 | < 2.0 |
| MQ | < 40.0 | - |
| FS | > 60.0 | > 200.0 |
| SOR | > 3.0 | > 10.0 |
| MQRankSum | < -12.5 | - |
| ReadPosRankSum | < -8.0 | < -20.0 |
| InbreedingCoeff | - | < -0.8 |

The unfiltered variants where exported to a table including the following information: Chromosome, position, reference allele, alternative allele, type of variant and reference allele frequency. This procedure was performed through the GATK tool *VariantsToTable*.



**Figure 2:** Workflow applied for obtaining significant variants which could be causative of obesity. The variant calling part in based on best practices workflow described by DePristo et.al. 201122 and Van der Auwera et.al 201321. The complete code of this workflow is stored in the following GitHub repository: https://bit.ly/2ELJ6R4

### Statistical analysis

In order to identify the significant variants, the minor allele frequency of each annotation was tested by comparing it with the minor allele frequency of European population from 1000 genomes project 23. By this way, those variants common in the samples but not common in the population where samples belong to can be identified. The procedure for obtaining these variants are explained as follows.

* **Minor Allele Frequency (MAF) calculation.** For calculating the MAF data, the refined VCF file obtained from variant calling procedures was loaded through the R package *vcfR*24. This package allowed to calculate the MAF of each annotation and the number of individuals who had the variant. The MAF information and number of individuals were included into the variant table. In addition, the minor allele frequency from European population from 1000 genomes project was also added to the variant table*.* A subset of the variant table is exposed in Table 3.

**Table 3**: Subset of 4 annotation from the variant table after including the minor allele frequency from 1000 genomes project (EUR\_MAF) and the number of individuals from Europe used to obtain this MAF (N\_eur). N\_ob is the number of people in the samples who contained the variant.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chromosome | Position | MAF | EUR\_MAF | N\_ob | N\_eur |
| 1 | 494515 | 0.083 | 0.020 | 6 | 669 |
| 1 | 591452 | 0.071 | 0.020 | 7 | 669 |
| 1 | 591460 | 0.071 | 0.030 | 7 | 669 |
| 1 | 598934 | 0.167 | 0.000 | 6 | 669 |
| … | … | … | … | … | … |

* **Association analysis.** It is a methodology useful for discovering relationship hidden in large datasets25. Performing the association analysis was useful for discovering which annotation is significantly different evaluating its MAF (MAF obtained from obese samples) and the MAF that they should has (MAF from control European individuals [1000 genomes project26]). For doing that, the fisher exact test was applied. Fisher exact test is based on the hypergeometric distribution where, considering the population size and allele frequencies, the association probability (P) can be calculated as follows27:

Considering the subset exposed in Table 3, the size of population (N) is obtained by summing the number of obese samples (N\_ob) and the number of control samples (N\_EUR). Considering the population and frequencies, the number of people with and without the allele is calculated multiplying the number of individuals by the allele frequency. By this way, the following matrix is obtained:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Allele | Non-allele | Total |
| Obese | ***a***: N\_ob x MAF | ***b:*** N\_ob x (1 - MAF) | ***r1:*** N\_ob |
| Controls | ***c:*** N\_EUR x AF\_EUR | ***d:*** N\_EUR x (1-AF\_EUR) | ***r2:*** N\_EUR |
|  | ***c1:*** *a+c* | ***c2:*** *b+d* | ***N: r1+ r2*** |

Once this matrix is created, the association probability can be calculated by the following formula28:

In practice, a R function was written to create the matrix and to perform the fisher test iteratively (annotation by annotation).

* **Multiple testing, False Discovery Rate (FDR) correction**. Because a separate statistical test was performed at each locus, traditional p-value cutoff of 0.01 and 0.05 had to be more stricter to avoid and abundance of false positive results29. For this reason, a multiple testing applying the FDR correction28 was used. The FDR is the proportion of the rejected null hypotheses which are erroneously rejected28, therefore, applying the correction, a new p-value per annotation is calculated in order to decrease the variants erroneously considered as significant.

In practice, this part of the analysis was performed by applying the R function *p.adjust*  to the p-values,specifying “FDR” as a method.

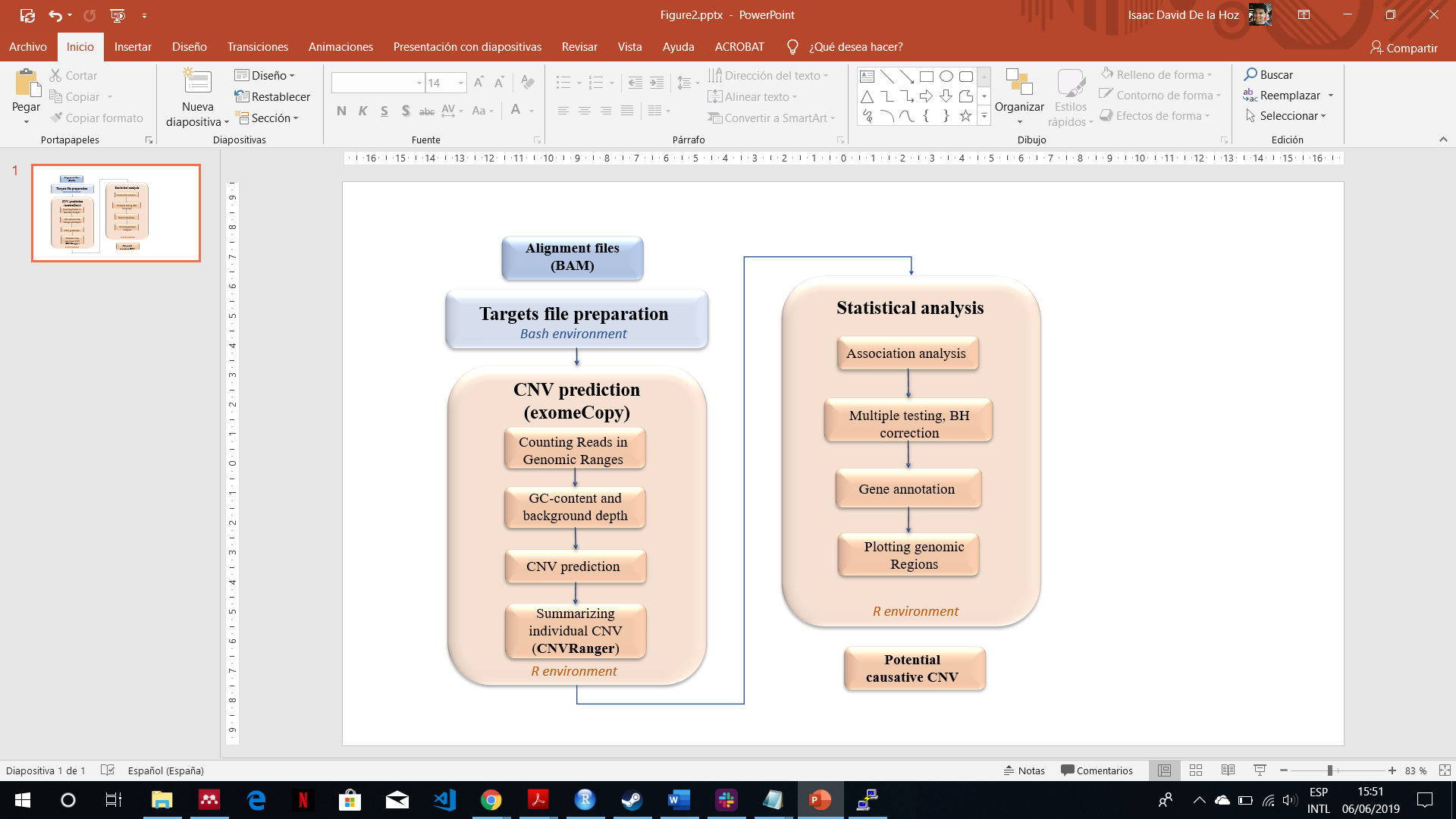
* **Pathway enrichment.** The enrichment analysis has the objective of interpreting gene expression based on functional annotation of the differentially expressed genes. The biochemical pathway, molecular function or biological process where these genes are involved are shown30.

In this case, the genes, where significant variants (p-value adjusted ≤ 0.05 and ≤ 0.01) were located, were annotated in order to perform the enrichment analysis over them via hypergeometric test using both Gene Ontology (GO)31 and Kyoto Encyclopedia of Genes and Genomes (KEGG)32 as a standardised annotation of gene products, with the objective of seeing if the molecular pathways of these genes are related to obesity.

* **Overlap with existing databases.** Finally, the MAF database from genomAD33 and Trans omics pression medicine (TopMed)34 projects were consulted in order to see both if these variants had already been recorded and the MAF reported in European population. This information was important because if the significant variants were not common in any population or were not already found, it would be very possible these variants were important in obesity.

## CNV Analysis

As the sequencing step is stochastic process, the number of times a sequence is read should be reflective of its relative copy number variants in the original sample. However, because of the hybridization step, many of the biases present in array-based studies, such as batch effects and, effects from obtaining DNA from different sources are still present in these data35. Nevertheless, the exome sequencing data can be used for finding CNVs which overlap exons and are not common in the control set36. So, in order to detect those exons a workflow was implemented (Figure 3). All parts of this workflows are explained in the following sections and the complete code can be found in this GitHub repository https://bit.ly/2K2qdxg.



**Figure 3**: workflow for obtaining significative CNV that could be related with obesity. The complete code is stored in the following repository: <https://bit.ly/2K2qdxg>.

### CNV prediction.

The prediction of CNVs was performed using the R package *exomeCopy*37. This package implements a hidden Markov model for predicting CNVs from exome sequencing experiments. This procedure was divided in the parts explained as follows.

* **Targets file preparation**. Due to the samples were exomes, the *exomeCopy* package needed a BED file containing the exon annotations. This information was taken from NCBI table browser38 and it was refined applying some simple Bash commands. By this way, non-relevant information was removed and the records were sorted. In addition, the annotation from chromosome X and Y were removed due to the further analysis was designed only for autosomal regions37.
* **Counting reads in genomic ranges**. Once the targets file was prepared, this package, through the function *countBamInRanges,* counted reads from the BAM files in genomic ranges covering the targeted regions. This function returned a vector of counts, representing the number of sequenced read starts (leftmost position regardless of strand) with mapping quality above a minimum threshold for each genomic range.
* **GC-content and background depth.** Besides counts reads, the GC-content and background depth were also needed to perform the CNV prediction. The GC-content was calculated by the function *getGCcontent* and the background depth by the function *generateBackground*. The first function calculated the GC-content using the targets file and the FASTA file of the reference genome. The last function applied 3 simple steps to calculate the background: [1] Given a vector of names of samples to be used as background, it extracted the read counts data frame from the Grange object, [2] it divided each sample by its mean read count (column means) and [3] it calculated the median of these normalized read counts (row medians).

The relationship between read counts and GC-content over the ranges varies across protocols and samples. It can be roughly approximated using second order polynomial terms of GC-content per sample, hence, a new column with the square of GC-content was added. In addition, a column with the width of the ranges was also added.

* **CNV prediction.** Once all information needed to perform the prediction was calculated (read counts, GC-content, background depth and range width) the *exomeCopy* prediction was performed. *exomeCopy* models the sample read counts on one chromosome as emitted observations of a hidden Markov model (HMM), where the **hidden state is the** **copy number of the sample**. The emission distributions are modelled with negative binominal distributions, as the reads counts from high-throughput sequencing are often overdispersed for the Poisson distribution. The following equation explains this model37:

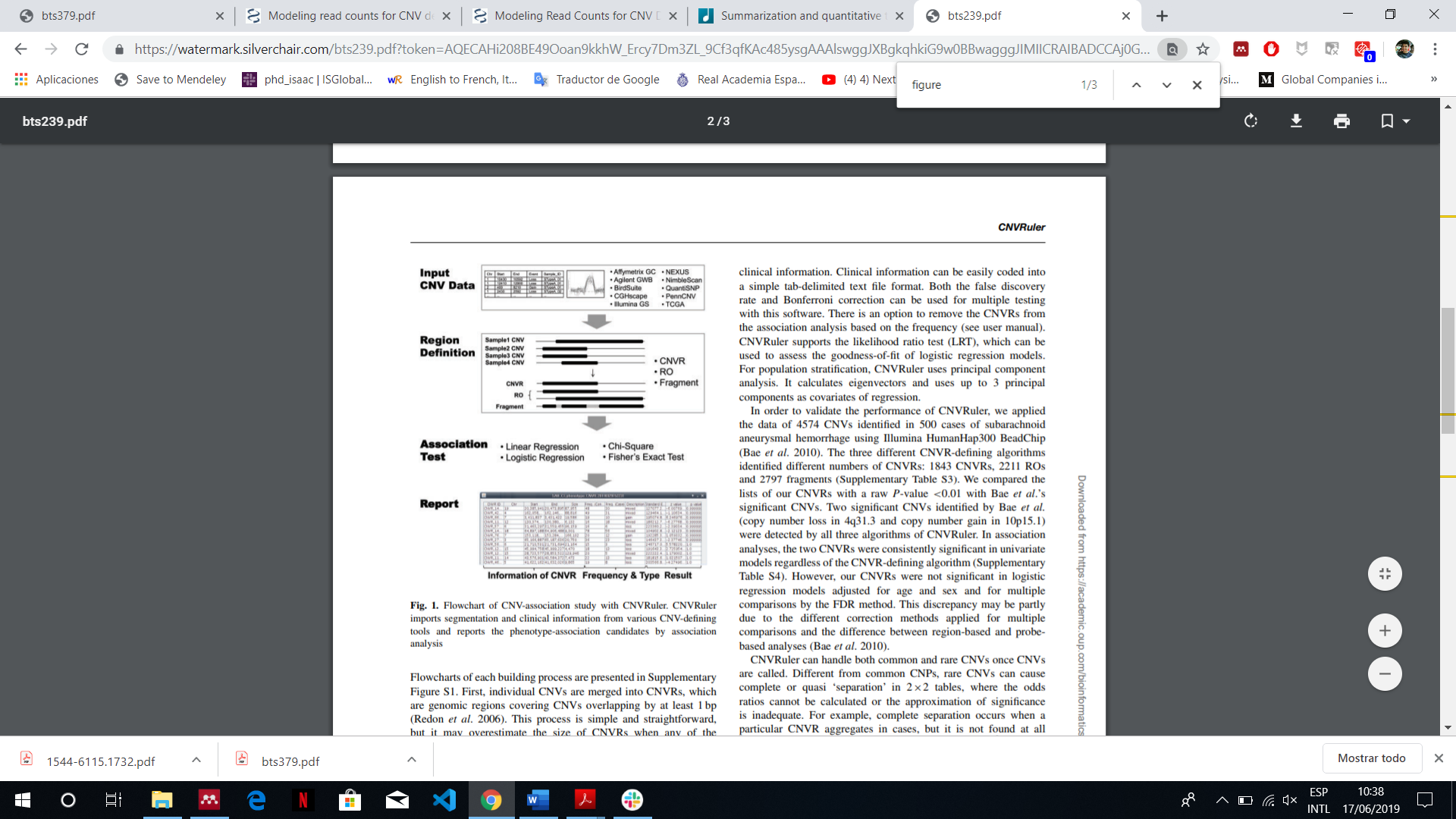
The mean parameter, , for genomic range *t* and hidden state *i* is a product of the possible copy number state over the expected copy number *d* (2 for diploid, 1 for haploid) and an estimate of the positional effects. The positional effects modelling comes from an exponential , where *xt* is the row of *X* (matrix of covariates) and *β* is a column of vector coefficients. The estimated positional effect is a combination of log background read depth, GC-content, range width, and any other covariate that is stored in matrix *X,* with a row for each range and a column for each covariate. The log of background read depth is used so that the counts and read depth come out on the same scale37.

The coefficients *β* are fit by the model to assess the likelihood of the HMM over all hidden state paths. In this way, the normalization and segmentation steps are combined into one step of maximizing the likelihood of the parameters given the data. The Viterbi algorithm is then applied to provide the most likely path37 producing, as a consequence, the more likely segments and their states.

The *exomeCopy* function creates a fitted object that need to be unfitted. For doing that, the function *compileCopyCountSegments* was executed on that fitted object generating a Grange object which provided the segments with the predicted copy number (state), the log odds of read counts being emitted from predicted copy count over normal copy count, the number of input genomic ranges containing within each ranges, the number of targeted basepairs contained in the segments, and the name of the sample to help compile the segments across the sample.

The expected state in autosomal chromosomes is 2, hence, the segments with a deviation either lower or higher than 2 can be considered as CNVs. For this reason, in order to have only CNV to work with, all segments with a state equal to 2 were removed.

* **Summarizing individual CNVs.** Once the CNV were predicted, the next step was **to merge overlapping individual calls** **into summarized regions** through the function from R package *CNVRanger*39 named *populationRanges.* This function apply the methodology from *CNVRuler*40 (Figure 4).



**Figure 4:** From CNV call from different samples the CNV ranges (CNVR), ranges determined by reciprocal overlap (RO) and simple overlapping fragments can be obtained. The arguments of the function *populationRanges* were set to obtain only the CNVR.

In addition, this function trimmed low-density areas (lower than 10%) of the number of calls within a summarized region. At the end, a Grange object containing the CNVs regions, the frequency of them among the samples was obtained and a column given information about what type the CNVs were, being “gain” when the region was overrepresented among the population, “loss” when the region was underrepresented and “both” when there was a combination of “gains” and “losses” of that region among the population.

All this procedure was also performed using **15 exomes of Iberian individuals** that were obtained from 1000 genomes projects in order to have **controls to compare with** in the further statistical analysis.

### Statistical analysis

For identifying the CNVs related with obesity, association analyses were performed over the frequencies of the overlapping CNVs predicted before in cases and controls (Figure 2). The regions significantly different were plotted and the genes affected were determined. The complete procedure is explained as follows.

* **Association analysis.** This procedure had the same aim that the applied in the section SNV detection and, as that section, the fisher test was the statistical method applied to discover which CNV, among those that overlap in case and controls, had a frequency significantly different.

In this case, the matrix constructed to perform the fisher test had the structure exposed in Table 4. From that, the fisher test analysis gave as output the probability (P) that the frequencies do not have significant differences between case and controls (null hypothesis). Hence, the CNVs with a P lower than 0.05 reject the null hypothesis, so they have a significant difference.

**Table 4:** Matrix structure created to perform the association analysis per each overlapping CNV. Freq-ob: frequency obesity,

|  |  |  |  |
| --- | --- | --- | --- |
|  | frequency | n. Individuals | Total |
| Obese | ***a***: Freq-ob | ***b:*** N\_ob | ***r1:*** N\_ob + Freq-ob |
| Controls | ***c:*** Freq-con | ***d:*** N\_con | ***r2:*** N\_con + Freq-con |
|  | ***c1:*** *a+c* | ***c2:*** *b+d* | ***N: r1+ r2*** |

* **Multiple testing, Benjamini Hochberg (BH) correction.** This procedure was applied with the same objective that the applied in SNV detection, to adjust the p-value in order to reduce the CNVs erroneously considered as significant. In this case, the BH correction, which is an alias of “FDR”, was applied. The CNVs with a p-value adjusted lower than 0.05 were selected.
* **Gene annotation and Plotting genomics regions.** The function *plotCNVs* from R package *gada*41, which is designed to visualize CNVs and genes in genomic regions, was modified to be able to work with our data, and, in this way, to visualize the genomic regions beside the genes and the frequencies of the overlapping CNVs significantly different between case and controls. The function *plotCNVs* uses the R package *Gviz*42 as engine to perform the visualization.

# Results

As all procedures took place, large datasets were produced as a result of the methodologies before explained. In order to expose the results, subsets of this datasets in form of tables were used. The complete tables are stored in the following repository: https://bit.ly/2RlBYA0

## SNV detection.

### Potential causative SNV

From variant calling of the obese exome sequencing data, a total of 2252592 genetic variants were found and exported to a variant table (Table 5).

**Table 5**: Subset of variant table obtained from variant calling procedure including MAF, number of samples, MAF of controls and number of controls.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| CHROM | POS | REF | ALT | TYPE |
| 1 | 19190 | GC | G | INDEL |
| 1 | 66169 | TA | T | INDEL |
| 1 | 98921 | AG | A | INDEL |
| 1 | 102951 | C | T | SNP |
| 1 | 132991 | G | A | SNP |
| 1 | 133129 | G | A | SNP |
| 1 | 133160 | G | A | SNP |
| … | … | … | … | … |

After converting this table in genomic range and including the MAFs and number of samples, the data had the configuration exposed in Table 3. From this data, the column containing the p-values was added after performing the fisher test and other columns were added after performing the multiple testing with FDR correction and gene annotation (Table 6).

**Table 6**: subset of the variant once the statistical analysis had been performed. The p-value was calculated through fisher test associating the number of obese and controls samples with the MAF of both. The p-values were then adjusted via multiple test with FDR correction.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chr | POS | MAF | N\_ob | Type | 1k\_MAF | N\_1k | P-value | P-val.adj. | GENE |
| 1 | 16875549 | 0.375 | 4 | INDEL | 0 | 669 | 2.65E-05 | 3.82E-02 | CROCC |
| 1 | 16876782 | 0.5 | 4 | SNP | 0 | 669 | 2.65E-05 | 3.82E-02 | CROCC |
| 1 | 16877558 | 0.429 | 7 | SNP | 0 | 669 | 6.83E-07 | 1.59E-03 | CROCC |
| 1 | 20997361 | 0.35 | 10 | INDEL | 0.005 | 669 | 8.20E-07 | 1.86E-03 | EIF4G3 |
| 1 | 21563291 | 0.233 | 15 | INDEL | 0 | 669 | 8.57E-06 | 1.44E-02 | ALPL |
| 1 | 22010317 | 0.458 | 12 | INDEL | 0 | 669 | 6.59E-10 | 5.71E-06 | CELA3A |
| 1 | 23967872 | 0.344 | 16 | SNP | 0.02 | 669 | 1.31E-06 | 2.86E-03 | SRSF10 |
| … | … | … | … | … | … | … | … | … | .. |

After comparing the MAFs of the obese people with the MAFs of the 669 Europeans controls in the 1000 Genomes Project, **704 SNVs involved in 479 different genes** were associated with obesity significantly at p-values **adjusted ≤ 0.05** of which **576 of them had a p-values adjusted ≤ 0.01** affecting a more than **386 genes** (Table S1 and S2 in repository).

### Enrichment analysis.

From GO enrichment, it was found that the genes affected by the significant variants covered **137 different gene functions**. The most significant were the following: [1] the platelet activation, [2] cellular response to nitrogen compound, [3] homocysteine metabolic process and sulphur amino acid metabolic process (Table 7).

**Table 7**: most significant gene functions after hypergeometric test comparing the expected counts of gene per function with the count found in the samples.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GOBPID | | Pvalue | | OddsRatio | | ExpCount | | Count | | Size | | Term | |
| GO:0030168 | 1.78E-05 | | 4.54 | | 3.17 | | 13 | | 145 | | platelet activation | |
| GO:1901699 | 1.42E-04 | | 2.33 | | 12.46 | | 27 | | 570 | | cellular response to nitrogen compound | |
| GO:0050667 | 1.89E-04 | | 18.11 | | 0.31 | | 4 | | 14 | | homocysteine metabolic process | |
| GO:0000096 | 2.46E-04 | | 7.79 | | 0.90 | | 6 | | 41 | | sulfur amino acid metabolic process | |
| GO:0071548 | 2.46E-04 | | 7.79 | | 0.90 | | 6 | | 41 | | response to dexamethasone | |
| GO:0060312 | 3.39E-04 | | 33.86 | | 0.15 | | 3 | | 7 | | regulation of blood vessel remodelling | |

From KEGG enrichment, it was found that the genes affected by the significant variants were part of **19 different metabolic pathways.** The most significant were the following: Focal adhesion, Dilated cardiomyopathy, FC gamma R-mediated phagocytosis and Amoebiasis (Table 8).

**Table 8**: most significant metabolic pathways determined by a hypergeometric test comparing the expected counts of genes per pathway with the count found in the samples.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| KEGGID | Pvalue | OddsRatio | ExpCount | Count | Size | Term |
| 4510 | 7.75E-06 | 4.229 | 4.474 | 16 | 196 | Focal adhesion |
| 5414 | 1.71E-04 | 5.120 | 2.031 | 9 | 89 | Dilated cardiomyopathy |
| 4666 | 2.03E-04 | 4.993 | 2.077 | 9 | 91 | Fc gamma R-mediated phagocytosis |
| 5146 | 4.81E-04 | 4.393 | 2.328 | 9 | 102 | Amoebiasis |
| 4810 | 7.61E-04 | 3.108 | 4.702 | 13 | 206 | Regulation of actin cytoskeleton |
| 4972 | 1.55E-03 | 4.048 | 2.214 | 8 | 97 | Pancreatic secretion |
| 4010 | 1.85E-03 | 2.680 | 5.820 | 14 | 255 | MAPK signaling pathway |
| 4512 | 2.17E-03 | 4.294 | 1.826 | 7 | 80 | ECM-receptor interaction |

### Variant selection and overlapping with existing databases.

The variants highly representative in samples (N\_ob ≥ 15, MAF ≥ 0.45) but rare in control samples (EU\_MAF ≤ 0.05) were selected obtaining the variants exposed in Table 10. **20 variants were found**.

After consulting with other MAF databases (genomAD33 and TopMed34) the list of significant variants (Table 10) was refined with those variants that presented a MAF ≈ 0 or NA (not detected) in these databases. **11 SNVs were found with these characteristics** (Table 9).

**Table 9**: Refined table of variants that are rare in in all control MAF databases (1000 G. project, TOPMED [TM] and GenoMad [GMAD]) and common in obese samples.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Chr. | Pos. | Ref/Alt | MAF | EU\_MAF | GENES | MAF\_TM | MAF\_GMAD |
| 1 | 3872630 | G/GCGGCCC | 0.5 | 0 | DFFB | 0.02 | NA |
| 1 | 85108022 | A/AC,ACT | 0.469 | 0.04 | WDR63 | 8.00E-06 | 0 |
| 1 | 150555614 | G/GACAC | 0.467 | 0.04 | ADAMTSL4 | 3.00E-04 | 1.00E-04 |
| 2 | 46356139 | C/CG | 0.5 | 0.001 | EPAS1 | NA | NA |
| 6 | 350940 | T/TA | 0.469 | 0.003 | DUSP22 | NA | NA |
| 6 | 38860648 | G/GT | 0.469 | 0.002 | DNAH8 | 6.00E-04 | 4.00E-04 |
| 7 | 21867834 | G/GT | 0.5 | 0.001 | DNAH11 | NA | NA |
| 12 | 865142 | T/TC | 0.5 | 0.005 | WNK1 | 0.003 | 0.004 |
| 14 | 104945729 | C/T | 0.5 | 0 | AHNAK2 | 3.00E-07 | 0.41 |
| 17 | 42702516 | G/GT | 0.469 | 0 | EZH1 | 2.00E-05 | 0 |
| 20 | 45419425 | C/CG | 0.469 | 0 | PIGT | 7.00E-13 | 2.00E-04 |

**Table 10:** variants that were very common in obese samples but rare in controls from 1000 genomes project. The variants highlighted are the common in controls from TopMed (65000 samples) and/or GenMad (5752 samples from southern Europe).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chr. | Pos. | Ref/Alt | MAF | EUR\_MAF | N\_ob | Padj.fdr | GENES | MAF\_TM | MAF\_GMAD |
| 1 | 3872630 | G/GCGGCCC | 0.5 | 0 | 15 | 9.82E-10 | DFFB | 0.02 | NA |
| 1 | 85108022 | A/AC,ACT | 0.469 | 0.04 | 16 | 5.47E-04 | WDR63 | 8.00E-06 | 0 |
| 1 | 150555614 | G/GACAC | 0.467 | 0.04 | 15 | 3.94E-03 | ADAMTSL4 | 3.00E-04 | 1.00E-04 |
| 2 | 46356139 | C/CG | 0.5 | 0.001 | 16 | 6.92E-09 | EPAS1 | NA | NA |
| 6 | 350940 | T/TA | 0.469 | 0.003 | 16 | 2.93E-08 | DUSP22 | NA | NA |
| 6 | 38860648 | G/GT | 0.469 | 0.002 | 16 | 6.92E-09 | DNAH8 | 6.00E-04 | 4.00E-04 |
| 7 | 21867834 | G/GT | 0.5 | 0.001 | 16 | 6.92E-09 | DNAH11 | NA | NA |
| 10 | **46550016** | C/T | 0.5 | 0.01 | 16 | 8.84E-07 | GPRIN2 | **0.5** | 0 |
| 12 | 865142 | T/TC | 0.5 | 0.005 | 16 | 8.52E-08 | WNK1 | 0.003 | 0.004 |
| 14 | **104945729** | C/T | 0.5 | 0 | 15 | 9.82E-10 | AHNAK2 | 3.00E-07 | **0.41** |
| 17 | **21298539** | A/G | 0.5 | 0 | 16 | 9.82E-10 | MAP2K3 | **0.5** | NA |
| 17 | **21298582** | C/T | 0.5 | 0 | 16 | 9.82E-10 | MAP2K3 | **0.5** | NA |
| 17 | **21298622** | T/C | 0.467 | 0.04 | 15 | 3.94E-03 | MAP2K3 | **0.5** | NA |
| 17 | **21298879** | C/A | 0.469 | 0 | 16 | 9.82E-10 | MAP2K3 | **0.5** | **0.5** |
| 17 | **21298925** | G/C | 0.469 | 0.001 | 16 | 6.92E-09 | MAP2K3 | 9.00E-11 | **0.5** |
| 17 | **21298960** | C/G | 0.469 | 0 | 16 | 9.82E-10 | MAP2K3 | **0.5** | **0.5** |
| 17 | **21303304** | G/A | 0.469 | 0 | 16 | 9.82E-10 | MAP2K3 | **0.5** | NA |
| 17 | 42702516 | G/GT | 0.469 | 0 | 16 | 9.82E-10 | EZH1 | 2.00E-05 | 0 |
| 20 | 45419425 | C/CG | 0.469 | 0 | 16 | 9.82E-10 | PIGT | 7.00E-13 | 2.00E-04 |

## CNV analysis

### Counts of reads.

Once the counting procedure of the reads was performed over the BAM files, a genomic range (Grange) object containing the number of reads per exons per sample was obtained. In addition, the GC-content and background depth, the square of GC-content and the annotation width were calculated over the counts obtaining an object with all in information needed to perform the *exomeCopy* model (this object is exposed briefly in Table 11).

**Table 11**: subset of the object used to perform the exome copy model. The information about counts is exposed in the sample name’s columns.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chr. | Start. | End | width | Sample.1 | … | Sample.15 | GC | GC.sq | bg |
| 1 | 12975 | 13052 | 78 | 20 | … | 21 | 0.6026 | 0.3631 | 0.0071 |
| 1 | 13221 | 13374 | 154 | 126 | … | 169 | 0.5909 | 0.3492 | 0.0655 |
| 1 | 13221 | 14409 | 1189 | 221 | … | 365 | 0.5627 | 0.3166 | 0.1162 |
| 1 | 13453 | 13670 | 218 | 50 | … | 117 | 0.5826 | 0.3394 | 0.0335 |
| 1 | 15796 | 15947 | 152 | 0 | … | 22 | 0.6316 | 0.3989 | 0.0066 |
| … | … | … | … | … | … | … | … | … | … |

### ExomeCopy model

Once all data is ready, the mode was applied per chromosome per sample through a wrapper function. This procedure gave as a result a fit object that once unfitted a new object was created containing **392593 predicted** **segments (549742 in controls)** and, among other information (Table 12), its copy count (the copy number of the segment).

**Table 12**: object resulting from *exomeCopy* model. This object contains each segment predicted, its copy count (hidden state), the log odds ratio associated to this copy count, the number of input genomic ranges containing within each range, and the name of the sample. The segments with a copy count different to 2 (expected normal state) can be considered CNVs. Hence, only retrieving the segments with a copy count = 2, the CNVs are obtained.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Chr. | start | end | width | copy.count | log.odds | nranges | sample.name |
| 1 | 6135230 | 6186816 | 51587 | 4 | 12.61 | 31 | Sample.1 |
| 1 | 6192927 | 6281286 | 88360 | 2 | 0 | 117 | Sample.1 |
| 1 | 6281241 | 6282888 | 1648 | 0 | 8.38 | 8 | Sample.1 |
| 1 | 6294864 | 6385813 | 90950 | 1 | 3.68 | 21 | Sample.1 |
| 1 | 6385664 | 6413282 | 27619 | 0 | 19.09 | 18 | Sample.1 |
| … | … | … | … | … | … | … | … |

### Summarizing individual CNVs (CNV Ranger).

Once obtained the CNVs predicted and summarized individual calls across the population, a Grange object was obtained containing, as ranges, the information about CNV regions formed by the CNV calls predicted by *exomeCopy* and, as metadata, the frequency of this region in the samples and the type (Table 13). A total of **1457 regions were obtained in samples (Table 13) and 10763 regions in controls** (Table S3 in repository).

**Table 13**: subset of CNV regions obtained from CNV calls in samples.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chr. | Start | End | Width | Freq | Type |
| 2 | 61090434 | 61108349 | 17916 | 2 | gain |
| 2 | 69963462 | 69963500 | 39 | 11 | both |
| 2 | 69996858 | 69996888 | 31 | 14 | both |
| 2 | 70086124 | 70086245 | 122 | 15 | loss |
| 2 | 86873840 | 86893375 | 19536 | 12 | both |
| … | … | … | … | … | … |

### Association analysis.

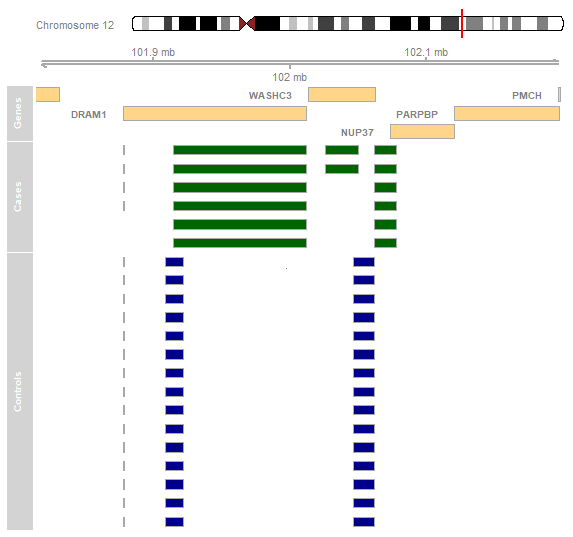
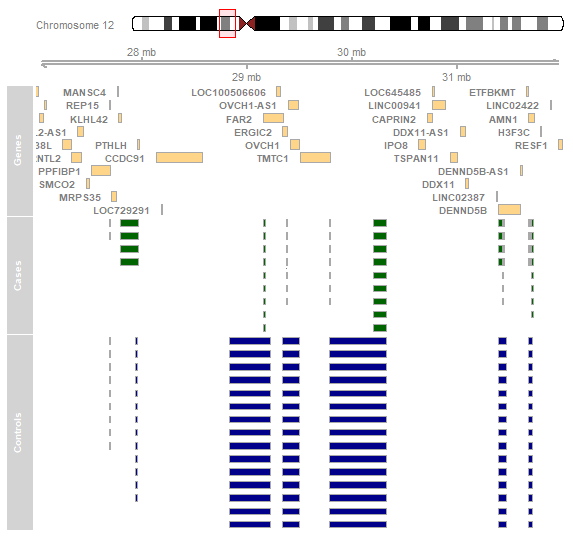
From summarized CNV regions from case and controls, it was found that **154 regions were shared between both**. This region was selected and it was applied the fisher test in order to determine the region with frequencies significantly different. It was found that from these 154 regions, **88 had a p-value ≤ 0.05 and 73 a p-value adjusted (BH correction) ≤ 0.05 regions** (Table 14)**.**

**Table 14**: Subset of table with the CNV regions with the significant frequency difference found after association analysis.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Chr. | start | end | width | freq | type | freqCtrl | typeCtrl | p.values | padj |
| 2 | 69996858 | 69996888 | 31 | 14 | both | 5 | loss | 1.70E-03 | 5.45E-03 |
| 2 | 70086124 | 70086245 | 122 | 15 | loss | 7 | both | 2.20E-03 | 6.27E-03 |
| 12 | 309803 | 309964 | 162 | 4 | both | 15 | both | 5.00E-05 | 2.57E-04 |
| 12 | 320995 | 323206 | 2212 | 4 | both | 15 | both | 5.00E-05 | 2.57E-04 |
| 12 | 323595 | 334422 | 10828 | 5 | both | 15 | both | 2.00E-04 | 8.55E-04 |
| 12 | 350621 | 385974 | 35354 | 5 | both | 15 | both | 2.00E-04 | 8.55E-04 |
| 12 | 389348 | 409421 | 20074 | 6 | both | 14 | both | 5.20E-03 | 1.38E-02 |
| … | … | … | … | … | … | … | … | … | … |

Almost all CNV regions were found in chromosome 12, so, the genomic regions which accumulated the highest amounts of CNVs where visualized giving as a result the figure 5 and 6.

**Figure 5**: genome visualization of the regions Chr12:27000000-32000000 [**A**] and Chr12:101814174-102200000 [**B**]. The frequency of CNVs regions among Cases (obese samples) represented as green boxes and Controls represented as blue boxes are shown.



**A**

**B**

# Discussion

# Conclusion

# Bibliography

1. Klaauw AA Van Der, Farooqi IS. Review The Hunger Genes : Pathways to Obesity. *Cell*. 2015;161(1):119-132. doi:10.1016/j.cell.2015.03.008

2. Moon S, Hwang MY, Jang HB, et al. Whole-exome sequencing study reveals common copy number variants in protocadherin genes associated with childhood obesity in Koreans. *Int J Obes*. 2017;41(4):660-663. doi:10.1038/ijo.2017.12

3. Adhanom Ghebreyesus T. WHO | What are the health consequences of being overweight? WHO. https://www.who.int/features/qa/49/en/. Published 2013. Accessed May 21, 2019.

4. Summers JB, Kaminski JM. Nutrition, physical activity, and obesity. *Lancet*. 2002;360(9341):1249. doi:10.1016/S0140-6736(02)11249-9

5. Tappy L, Binnert C, Schneiter P. Energy expenditure, physical activity and body-weight control. *Proc Nutr Soc*. 2004;62(03):663-666. doi:10.1079/pns2003280

6. Schwartz MW, Woods SC, Seeley RJ, Barsh GS, Baskin DG, Leibel RL. Is the Energy Homeostasis System Inherently Biased Toward Weight Gain ? 2003;52(February).

7. Wu Y, Wang W, Jiang W, Yao J, Zhang D. An investigation of obesity susceptibility genes in Northern Han Chinese by targeted resequencing. *Med (United States)*. 2017;96(7):1-6. doi:10.1097/MD.0000000000006117

8. Serra-Juhé C, Martos-Moreno G, Bou de Pieri F, et al. Novel genes involved in severe early-onset obesity revealed by rare copy number and sequence variants. *PLoS Genet*. 2017;13(5):1-19. doi:10.1371/journal.pgen.1006657

9. Yanovski JA. Obesity: Trends in underweight and obesity — scale of the problem. *Nat Rev Endocrinol*. 2017;14(1):5-6. doi:10.1038/nrendo.2017.157

10. Stunkard AJ, Harris JR, Pedersen NL, McClearn GE. The Body-Mass Index of Twins Who Have Been Reared Apart. *N Engl J Med*. 1990;322(21):1483-1487. doi:10.1056/NEJM199005243222102

11. Maes HHM, Neale MC, Eaves LJ. Genetic and Environmental Factors in Relative Body Weight and Human Adiposity. *Behav Genet*. 1997;27(4).

12. Speakman JR. Commentary A Nonadaptive Scenario Explaining the Genetic Predisposition to Obesity : The ‘“ Predation Release ”’ Hypothesis. 2007;(July):5-12. doi:10.1016/j.cmet.2007.06.004

13. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature*. 2015;518(7538):197-206. doi:10.1038/nature14177

14. Pettersson M, Viljakainen H, Loid P, et al. Copy Number Variants Are Enriched in Individuals With Early-Onset Obesity and Highlight Novel Pathogenic Pathways. *J Clin Endocrinol Metab*. 2017;102(8):3029-3039. doi:10.1210/jc.2017-00565

15. Thorleifsson G, Walters GB, Gudbjartsson DF, et al. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. 2009;41(1):18-24. doi:10.1038/ng.274

16. Walters RG, Jacquemont S, Valsesia A, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11.2. *Nature*. 2010;463(7281):671-675. doi:10.1038/nature08727

17. Bochukova EG, Huang N, Keogh J, et al. Large , rare chromosomal deletions associated with severe early-onset obesity. *Nature*. 2010;463(7281):666-670. doi:10.1038/nature08689

18. Yang T-L, Guo Y, Shen H, et al. Copy Number Variation on Chromosome 10q26.3 for Obesity Identified by a Genome-Wide Study. *J Clin Endocrinol Metab*. 2013;98(1):E191-E195. doi:10.1210/jc.2012-2751

19. Jarick I, Vogel CIG, Scherag S, et al. Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis. *Hum Mol Genet*. 2011;20(4):840-852. doi:10.1093/hmg/ddq518

20. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110

21. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In: *Current Protocols in Bioinformatics*. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2013:11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

22. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43(5):491-498. doi:10.1038/ng.806

23. Auton A, Abecasis GR, Altshuler (Co-Chair) DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

24. Knaus BJ, Grünwald NJ. VCFR : a package to manipulate and visualize variant call format data in R. *Mol Ecol Resour*. 2017;17(1):44-53. doi:10.1111/1755-0998.12549

25. Tan P-N, Steinbach M, Kumar V. Association Analysis: Basic Concepts and Algorithms. In: *Introduction to Data Mining*. Vol 19. ; 2006:88. doi:10.1111/j.1600-0765.2011.01426.x

26. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

27. Hoffman JIE. Hypergeometric Distribution. In: *Biostatistics for Medical and Biomedical Practitioners*. Elsevier; 2015:179-182. doi:10.1016/B978-0-12-802387-7.00013-5

28. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B*. 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x

29. Storey JD, Tibshirani R. Statistical significance for genomewide studies. *PNAS*. 2003;100(16):9440-9445.

30. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 2019;47(D1):D590-D595. doi:10.1093/nar/gky962

31. Mi H, Huang X, Muruganujan A, et al. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res*. 2017;45(D1):D183-D189. doi:10.1093/nar/gkw1138

32. Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27

33. Karczewski KJ, Francioli LC, Tiao G, et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv*. 2019:531210. doi:doi.org/10.1101/531210

34. Taliun D, Harris DN, Kessler MD, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *bioRxiv*. 2019. doi:10.1101/563866

35. Coin LJM, Cao D, Ren J, et al. An exome sequencing pipeline for identifying and genotyping common CNVs associated with disease with application to psoriasis. *Bioinformatics*. 2012;28(18):370-374. doi:10.1093/bioinformatics/bts379

36. Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, Haas SA. Modeling Read Counts for CNV Detection in Exome Sequencing Data. *Stat Appl Genet Mol Biol*. 2011;10(1):1-28. doi:10.2202/1544-6115.1732

37. Love M. Copy number variant detection in exome sequencing data using exomeCopy. *October*. 2012:1-15. http://bioconductor.org/packages/release/bioc/vignettes/exomeCopy/inst/doc/exomeCopy.pdf.

38. Karolchik D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*. 2004;32(90001):493D - 496. doi:10.1093/nar/gkh103

39. Geistlinger L, da Silva VH, Ramos M, Waldron L. CNVRanger: Summarization and expression/phenotype association of CNV ranges. 2019.

40. Kim J-H, Hu H-J, Yim S-H, Bae JS, Kim S-Y, Chung Y-J. CNVRuler. *Bioinformatics*. 2012;28(13):1790-1792. doi:10.1093/bioinformatics/bts239

41. Gonzalez JR, Pique-Regi R, Caceres A. gada: Genome Alteration Detection Algorithm (GADA). 2018. http://brge.isglobal.org.

42. Hahne F, Ivanek R. Visualizing Genomic Data Using Gviz and Bioconductor. In: Mathé E, Davis S, eds. *Statistical Genomics*. Vol 1418. Methods in Molecular Biology. New York, NY: Springer New York; 2016:335-351. doi:10.1007/978-1-4939-3578-9\_16