

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](#_Toc10556401)

[Approval and signature 4](#_Toc10556402)

[Abstract 5](#_Toc10556403)

[Table of contents 6](#_Toc10556404)

[1. Introduction 7](#_Toc10556405)

[1.1. Obesity 7](#_Toc10556406)

[1.1.1. Obesity in Spanish population 7](#_Toc10556407)

[1.2. Body weight control and causes of obesity 7](#_Toc10556408)

[1.2.1. Body weight control 7](#_Toc10556409)

[1.2.2. Pathway of energy homeostasis 8](#_Toc10556410)

[1.2.3. Disorder of energy homeostasis 9](#_Toc10556411)

[1.3. Importance of genetics in obesity 9](#_Toc10556412)

[1.4. Obesity susceptibility variants 9](#_Toc10556413)

[2. Objectives 11](#_Toc10556414)

[3. Material and Methods 12](#_Toc10556415)

[3.1. Data description 12](#_Toc10556416)

[3.1.1. Samples 12](#_Toc10556417)

[3.1.2. Methodology of DNA extraction and sequencing 13](#_Toc10556418)

[3.1.3. Genomic alignment 13](#_Toc10556419)

[3.2. SNV detection 13](#_Toc10556420)

[3.2.1. Variant Calling and annotation (GATK) 13](#_Toc10556421)

[3.2.2. Statistical analysis 15](#_Toc10556422)

[3.3. CNV Analysis 17](#_Toc10556423)

[4. Results 18](#_Toc10556424)

[5. Discussion 19](#_Toc10556425)

[6. Conclusion 20](#_Toc10556426)

[7. Bibliography 21](#_Toc10556427)

# 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

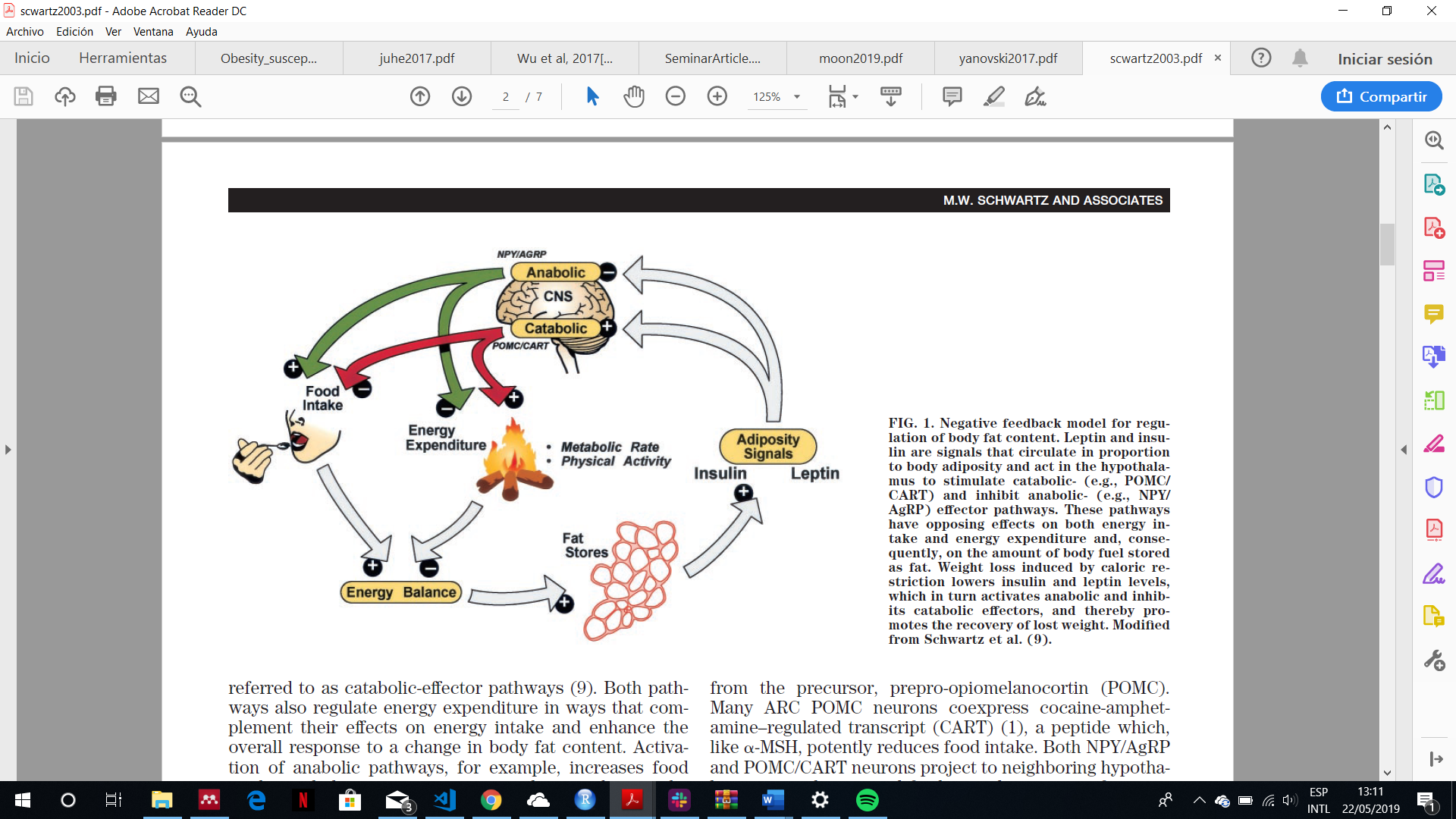
### Body weight control

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)9,14. 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 develop two workflows capable of analysing exome sequencing data for obtaining single nucleotide variants and copy number variants, respectively.
* To develop appropriate statistical analyses in order to discover new single nucleotide variants (such as SNPs or INDELs) and copy number variants highly correlated with obesity in Spanish population.
* 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 increased.

# 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 in order to find variants significantly associated with obesity (Figure 1). 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 entirely 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.

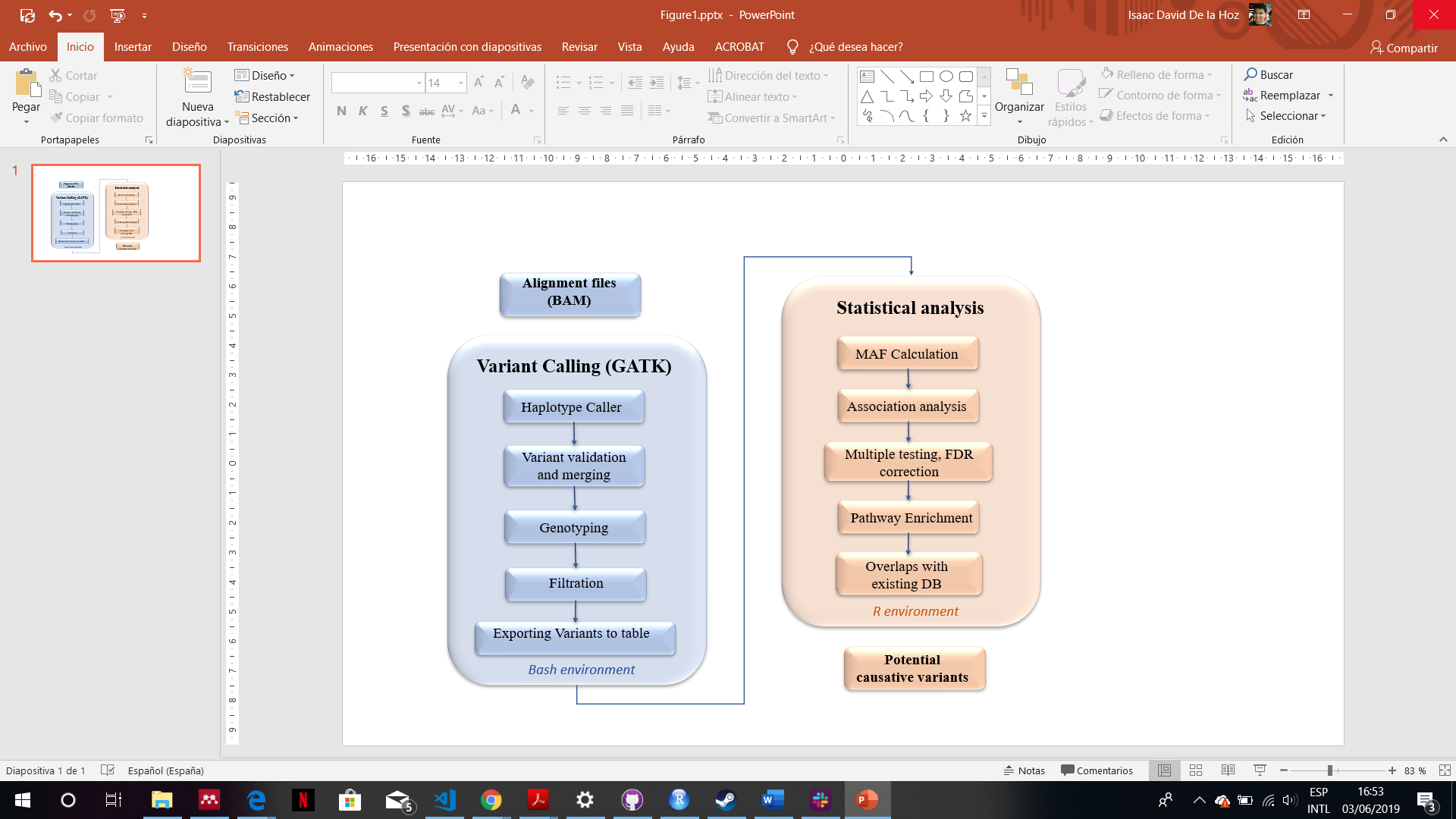
For the later statistical analysis was needed to calculate de minor allele frequency of each allele, and, for doing that, all the variants called from all samples had to be allocated in a single multi-VCF file. Therefore, according the best practices protocol21, the *haplotype caller* was run to every 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 the existence of any format error would make the merging step fail. 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 1:** 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

For identifying the significant variants, the minor allele frequency of each annotation was tested by comparing it with the allele frequency obtained 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*. 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\_AF) and the number of individuals from Europe used to obtain this AF.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Chromosome | Position | MAF | EUR\_AF | 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 datasets24. In our data, 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 project25]). 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 follows26:

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 formula:

In practice, a R function was written to create the matrix and 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 results27. 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 shown29.

In this case, the genes where significant variants (p-value adjusted ≤ 0.05) are located were annotated in order to perform the enrichment analysis over them via hypergeometric test using both Gene Ontology (GO)30 and Kyoto Encyclopedia of Genes and Genomes (KEGG)31 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 genomAD32 and TopMed33 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

Another workflow was developed in order to

# Results

# 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. 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

9. 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

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. 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

25. 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

26. 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

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

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. 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

30. 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

31. 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

32. 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

33. 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