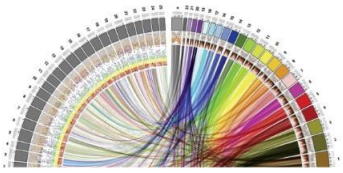

MSc in Bioinformatics

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

Universitat Autònoma de Barcelona

ISGlobal Institut de Salut Global Barcelona

July 2019

“I don't see the logic of rejecting data just because they seem incredible.”

Fred Hoyle

Acknowledgments

I would like to acknowledge to my supervisor, Juan Ramón González, for guiding me in the correct way during all this process, for being available at any time to any query, for correcting the master thesis once written and, finally, for treating me as colleague within his group. These things made me grow a lot both academically and professionally.

My sincere thanks to Raquel Rodríguez López and David dos Santos Albuquerque from The Genomics group of the Research Foundation of the General Hospital of Valencia (FIHGUV) for providing us the sequencing data of the individuals analysed in this master thesis.

In addition, I would like to thanks to all my workmates in ISGlobal for helping me whenever I had a programming query and for giving me support when I needed. They made an environment inside the company that made me enjoy working there.

Finally, but not less important, I want to thank to my family for having supported me every single day since I am in Barcelona. Without them I would have not be able to fulfil my desire of studying Bioinformatics. Thank you, family, I love you all.

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

Isaac David De la Hoz Saltaren

Student

UAB

Raquel Egea

Academic tutor

UAB

Juan Ramón González Ruiz

Thesis director

ISGlobal

Abstract

Obesity is a medical condition that is defined as excessive accumulation of fat that is sufficient to adversely affect health. The 10-20% of European population are classified as obese, being more severe in Spanish population where the 26.6% of adults are considered as obese and 62% are overweight. The obesity is commonly associated with environmental factor but also there are a genetic predisposition to a person become obese. Actually, in any environment either energy rich or energy lacking there are a considerable distribution of different body weight among people.

More than 97 genetic loci associated with obesity have been discovered where are located a big number of genes associated with the energy homeostasis such as LEP, LEPR, POMC and MC4R. In addition, single nucleotide variants (SNV) in genes such as FTO have been unequivocally associated with obesity in both childhood and adult obesity populations. On the other hand, several copy number variants (CNV) that contribute with the obesity heritability have been reported including deletions upstream of the NEGR1 gene, and distal deletions at 16p11.2, gains at 10q26.6 containing the CYP2E1 gene, among others.

From exome sequencing data of extremely obese individuals (BMI= 53.2 ±10.2), workflows were designed in order to, on the one hand, perform the variant calling and CNV prediction and, on the other hand, to apply the proper statistical analysis to associate those SNV and CNV with obesity. From these workflows, 704 SNVs involved in 479 different genes were significantly associated with obesity with p-values FDR-adjusted ≤ 0.05 and 73 CNV regions with p-value BH-adjusted ≤ 0.05.

Some molecular functions and biological pathway such as phospholipid metabolic process, platelet activation and focal adhesion, which are associated with obesity, were found after performing the enrichment analysis over the 479 genes affected by the 704 SNV. In addition, 329 of these genes had already been associated with obesity in previous GWAS studies, suggesting 150 new genes that could be associated too. On the other hand, from the 73 significant CNV region, only 26 were located in loci which had already been related with obesity. The other 47 CNV regions could suggest that new obesity-related loci.

In any case for both SNV detection and CNV analysis, extra studies have to be carried out to confirm or to reject the findings in this work. In addition, a larger number of samples would be preferable in subsequent analyses.

Table of contents

1. Introduction

1.1. Obesity

Obesity is a medical condition that is defined as excessive accumulation of fat that is sufficient to adversely affect health[1,2]. According World Health Organization (WHO), people with a body mass index (BMI; weight in kg/height in m^2) higher than 30 kg/m^2 are considered obese and higher than 25 kg/m^2 are considered overweight. The 30% of Americans and 10%–20% of Europeans are classified as obese, with the prevalence rising in many developing countries[1]. 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 disability[3].

1.1.1. 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 overweight[4].

1.2. Body weight control and causes of obesity

1.2.1. 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 expenditure[1]. 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 stores[5].

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 tissue[1,5].

1.2.2. Pathway of energy homeostasis

Pathways that stimulate food intake and promote weight gain are referred as anabolic-effector pathways, whereas those that promote anorexia and depletion of body fat are referred to as catabolic-effector pathways[6]. 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 pathways[6]. 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 other[6]. 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 neurons[6]. 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].

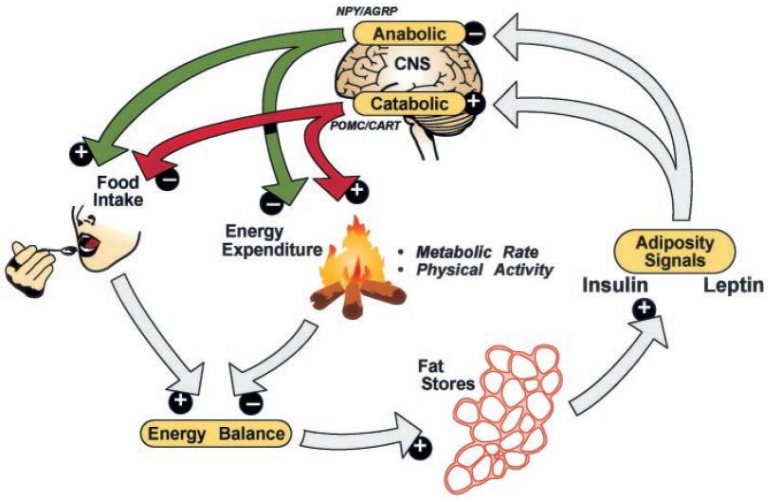

Figure 1: energy balance pathway. This model explains the body fat mass storage mechanism. When the food intake increases, the excess of energy is stored at adipocytes. Once the adipocytes have stored energy in form of fat, they produce leptin that activate central nervous system pathways which stimulate the decrease of energy intake (satiety signals) and the increase in energy expenditure. Figure adapted from Schwartz et.al 2003[6].

1.2.3. 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 for poor health worldwide[1].

Other factor that disrupt the energy homeostasis is genetics. 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 factors[1,2,7–9].

1.3. 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 children[10,11]. 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 comes[1,12].

1.4. Obesity susceptibility variants

As explained before, genetics factors have a big influence 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 factors[1]:

- 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) approaches[13]. 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 gene[15], proximal and distal deletions at 16p11.2[16,17], gains at 10q26.6 containing the CYP2E1 gene (MIM 124040)[18], and homozygous deletions at 11q11 encompassing olfactory receptor genes[19], among others.

In short, both single nucleotides variants (SNPs and Indels) and copy number variants may be focus of study not only for the increasingly high rate of discovery of variant related with obesity, but also because the variants already found only explain less than 3% of the heritability of BMI.

2. 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 find the biological significance of these variants in developing obesity analysing the genes and metabolic pathways affected.
- To find new genes and variants associated with obesity

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.

3. Material and Methods

3.1. Data description

3.1.1. Samples

The data come from 17 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. In addition, these individuals accomplished the following selection criteria:

- More than 3 Kg at be born
- To get the obesity being less than 6 years old
- BMI higher than 45 kg/m^2 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

Thanks to these selection criteria, the influence of external diseases that can cause obesity is avoided and, even more important if possible, we make sure that the samples have a great possibility of suffering obesity associated with their genetic component. These characteristics increase the power of finding susceptibility variants.

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/m^2)	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.*

3.1.2. 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/mm^2 on average.

3.1.3. 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.

3.2. 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).

3.2.1. Variant Calling and annotation (GATK)

The variant calling procedure was performed using GATK[20] tools following the best practices protocol recommended by GATK[21,22]. This procedure was divided in 4 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 regions[21].

  For the later statistical analysis, the minor allele frequency was needed to be calculated per each variant, and, for doing that, all the variants called from all samples had to be placed in a single multi-VCF file. Therefore, according the best practices protocol[21], the *haplotype caller* was run on every single sample with the **GVCF mode** activated to generate 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. In addition, additional information such as genotypes likelihoods and genotype quality were 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 through additional filters[21]. So, The VCF file was then filtered applying the recommended thresholds[21] exposed in Table 2.

Table 2: filtering parameters applied to the VCF file[21]. 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, 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 last procedure was performed through the GATK tool *VariantsToTable*.

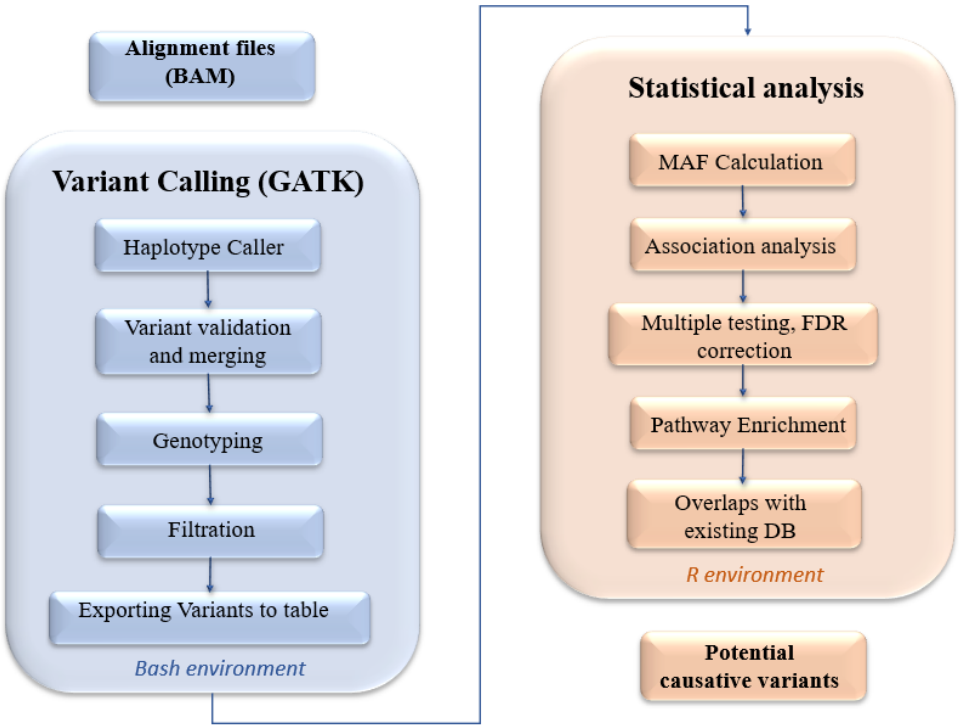

Figure 2: Workflow applied for obtaining significant variants which could be causative of obesity. The variant calling block is based on best practices workflow described by DePristo et.al. 2011[22] and Van der Auwera et.al 2013[21]. The complete code of this workflow is stored in the following GitHub repository: https://bit.ly/2ELJ6R4.

3.2.2. 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 into R 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 of

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 obese samples who had 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 datasets[25]. 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 project[26]]). 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[27]:

  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)	r_1**:** N_ob
CONTROLS	***c:*** N_EUR x AF_EUR	***d:*** N_EUR x (1-AF_EUR)	r_2**:** N_EUR
	c_1**:** *a*+*c*	***c2:*** *b*+*d*	***N:*** r_1**+** r_2

Once this matrix is created, the association probability can be calculated by the following formula[28]:

$$P = \frac{\binom{c_1}{a} * \binom{c_2}{b}}{\binom{N}{r_1}} = \frac{c_1!\, c_2!\, r_1!\, r_2!}{N!\, a!\, b!\, c!\, d!}$$

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 results[29]. For this reason, a multiple testing applying the FDR correction[28] was used. The FDR is the proportion of the rejected null hypotheses which are erroneously rejected[28], 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 determined[30].

  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.

- **Overlapping with existing databases and consulting with GWAS databases.** Finally, the MAF database from genomAD[33] 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 (MAF ≈ 0) or were not already found, it would be very possible these variants were new ones related with obesity.

  In addition, a genome wide association studies (GWAS) database[35] was consulted and genes associated with obesity were checked[36] in order to compare them with the genes associated with obesity found in this study.

3.3. 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 data[37]. Nevertheless, the exome sequencing data can be used for finding CNVs which overlap exons and are not common in the control set[38]. So, in order to detect those CNV 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.

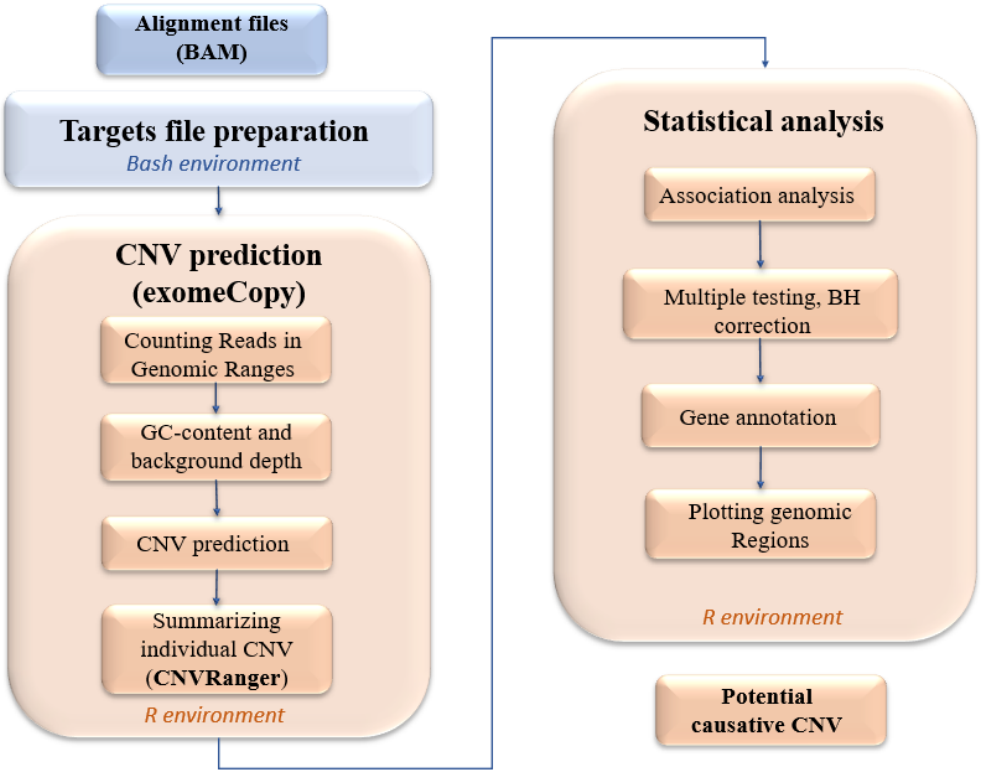

Figure 3: workflow for obtaining CNV regions that could be related with obesity. The complete code is stored in the following repository: https://bit.ly/2K2qdxg.

3.3.1. CNV prediction.

The prediction of CNVs was performed using the R package *exomeCopy*[39]. 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 browser[40] 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 regions[39].

- **Counting reads in genomic ranges**. Once the targets file was prepared, this package, through the function *countBamInRanges*, counted the reads of the BAM files in genomic ranges covering the targeted regions. This function returned a vector of counts, representing the number of sequenced reads start (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 model[39]:

$$\mu_{ti} = \frac{S_i}{d} e^{(x_t * \beta)}$$

  The mean parameter, μ_{ti}, for genomic range t and hidden state i is a product of the possible copy number state S_i 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 $(x_t * \beta)$, where x_t 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 scale[39].

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 path[39] 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 contained 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 contained within each segment, 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 calls were predicted, the next step was **to merge overlapping individual calls into summarized regions** through the function from R package *CNVRanger*[41] named *populationRanges.* This function apply the methodology from *CNVRuler*[42] (Figure 4).

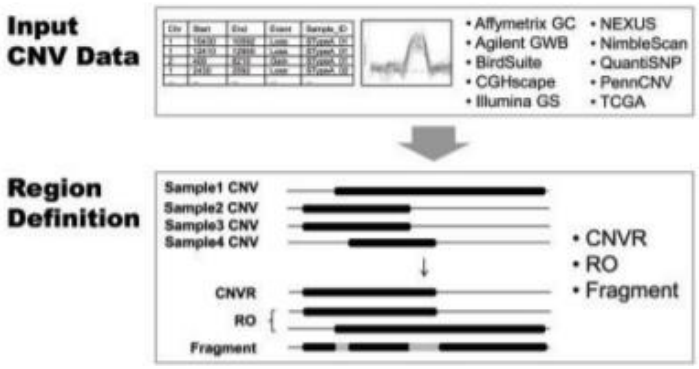

Figure 4: From the CNV calls obtained from different samples, the CNV ranges (CNVR), ranges determined by reciprocal overlap (RO) and simple overlapping fragments can be obtained. The arguments of the function *populationRanges* were set to obtain only the CNVR.

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 **CNV regions** and **the frequency of them among the samples** was obtained. In addition, the Grange object had a column containing the information about what **type the CNV regions** were, being "gain" when the region was formed by overrepresented CNV calls

(state > 2) among the population, "loss" when the region was formed by underrepresented CNV calls (state < 2) and "both" when there was a combination of "gains" and "losses" CNV calls in that region among the population.

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

3.3.2. Statistical analysis

For identifying the CNVs related with obesity, association analyses were performed over the frequencies of the overlapping CNV regions in cases and controls (Figure 2). By this way, the genomic loci where the significant CNV regions were located could be identified. 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 region, 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 association probability (P).

Table 4: Matrix structure created to perform the association analysis per each overlapping CNV. Freq-ob: frequency in obese samples, Freq-con: Frequency in control samples.

	FREQUENCY	**N. INDIVIDUALS**	**TOTAL**
OBESE	***a***: Freq-ob	***b:*** N_ob	r_1**:** N_ob + Freq-ob
CONTROLS	***c:*** Freq-con	***d:*** N_con	r_2**:** N_con + Freq-con
	c_1**:** $a+c$	***c2:*** $b+d$	***N:*** r_1+r_2

- **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*[43], 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*[44] as engine to perform the visualization.

4. Results

As a result of the methodologies before explained, larger datasets were produced and, in order to expose the results properly, subsets of this dataset in form of tables were presented. The complete tables are stored in the following repository: https://bit.ly/2RlBYA0

4.1. SNV detection.

4.1.1. 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.

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 object 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 was 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	826893	0.227	11	SNP	0.13	669	0.644269011	1	LINC01128
1	827209	0.15	10	SNP	0.13	669	0.627907902	1	LINC01128
1	827212	0.15	10	SNP	0.13	669	0.627907902	1	LINC01128
1	827221	0.15	10	SNP	0.13	669	0.627907902	1	LINC01128
1	827252	0.15	10	SNP	0.13	669	0.627907902	1	LINC01128
1	833068	0.125	4	SNP	0.09	669	1	1	LINC01128
1	833439	0.125	8	SNP	0.009	669	0.080186022	1	LINC01128
…	…	…	…	…	…	…	…	…	…

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 these SNVs had a p-values adjusted ≤ 0.01** affecting to **386 genes** (Table S1 and S2 in repository).

4.1.2. Enrichment analysis.

From GO enrichment, it was found that the genes affected by the significant variants covered **125 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 genes 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

4.1.3. Variant selection and overlapping with existing databases and comparison.

The variants highly representative in samples (N_ob ≥ 15, MAF ≥ 0.45) but rare in control samples (EU_MAF ≤ 0.05) were selected. **20 variants** were obtained (Table 10).

After consulting with other MAF databases (genomAD[33] and TopMed[34]) the list of significant variants (Table 10) was refined with those variants that presented a MAF ≈ 0 or NA (not detected) in these databases. **10 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
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

Finally, after comparing the 479 significant genes associated with obesity found in this study with obesity-related genes reported in GWAS studies, it was found that **329 of significant genes were into this database** suggesting that the other **150 genes and their variants** (TableS3 in repository) **are new ones associated with obesity.** If the variants with a MAF ≥ 0.05 in TOPMed project and/or GMAD project are removed, the number of associated genes is reduced until **122** (TableS4 in repository).

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

4.2. CNV analysis

4.2.1. 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
…	…	…	…	…	…	…	…	…	…

4.2.2. ExomeCopy model

Once all data is ready, the model 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 removing 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
…	…	…	…	…	…	…	…

4.2.3. 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 S5 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
…	…	…	…	…	…

4.2.4. Association analysis.

From summarized CNV regions from case and controls, it was found that there were **154 overlapping regions**. These regions were selected and the fisher test was applied 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 CNV regions with a frequency significantly different between case and control samples.

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. The genomic regions which accumulated the highest amounts of CNVs where visualized giving as a result the Figure 5A and 5B.

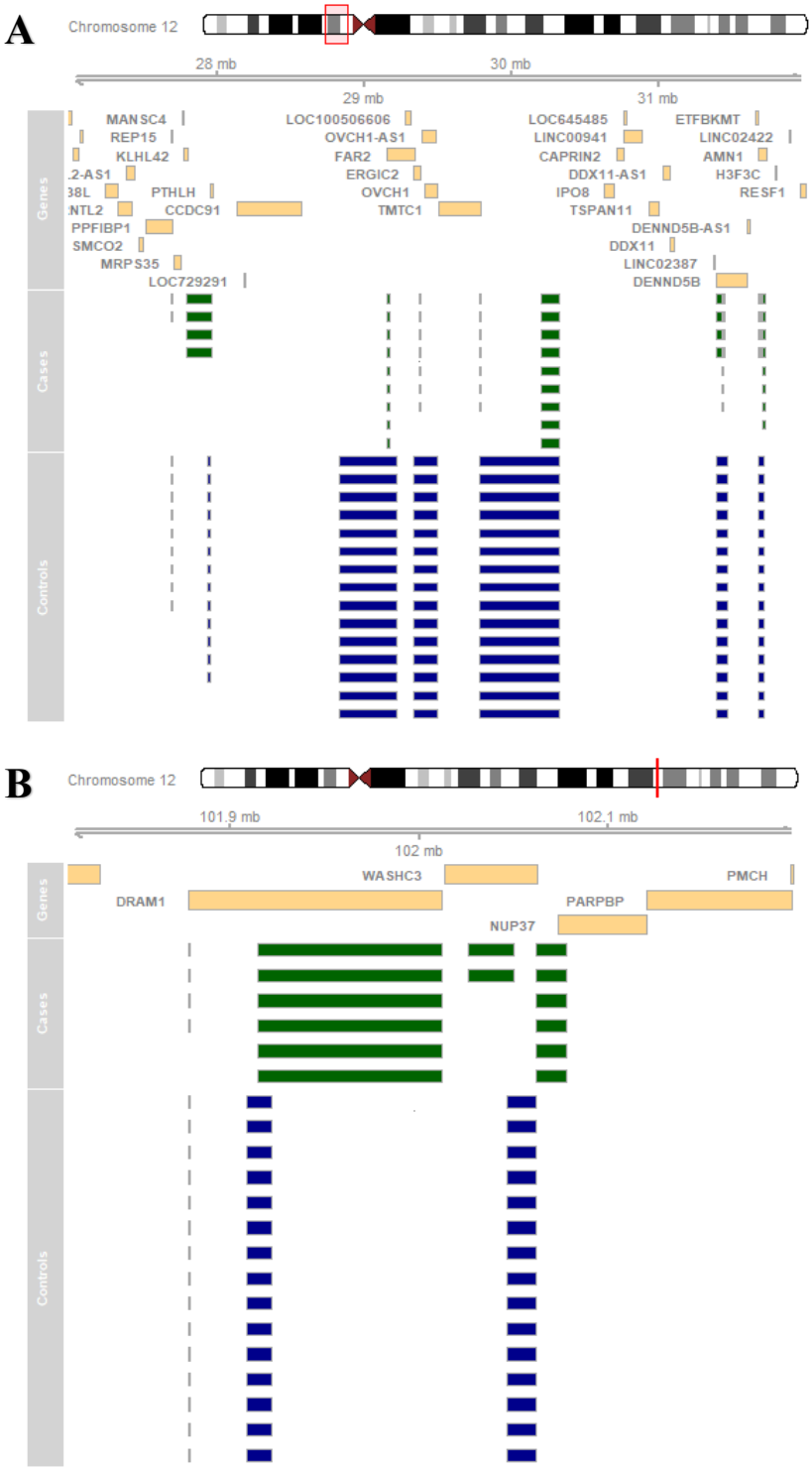

Figure 5: genome visualization of the regions Chr12:27000000-32000000 [**A**] and Chr12:101814174-102200000 [**B**]. The frequency of CNVs regions among Cases (obese samples) represented as green boxes and Controls represented as blue boxes are shown.

5. Discussion

From all procedures performed, a quite amount of data were resulted. This data allows us to face off it from different starting points. At SNV level, the enrichment results and the variant selection after overlapping with different databases, and, at CNV level, the loci where the significative CNV regions are located. All these points are analysed as follows.

5.1. Important SNVs discovered from SNV detection procedure.

5.1.1. Genes related with phospholipid metabolic process

From the procedure of SNV detection 704 significant SNV were detected (p-value [FDR] ≤ 0.05) affecting 479 genes (SNV-related genes). These genes were then analysed through GO enrichment in order to find some molecular function or metabolic pathways in which they are involved. It was found that one of the significant pathways where 17 of these genes (FLT1, INPP4A, LDLR, PDGFRA, PIK3CA, PRKCD, TGFB1, VAV2, EFR3A, PIP5K1C, PIK3R5, PLA2G15, RAB14, PIGT, ETNK2, FAM126A, TTC7B) are involved is the phospholipid metabolic process (p-value = 7.27E-04) (Table 15). In addition, the biological and functional importance of some of these genes in relation to obesity have already been reported.

The **FLT1** is the gene of the vascular endothelial growth factor receptor 1 (VEGFR1). This protein is important because it recognises vascular endothelial growth factor-1 (VEGF1) which controls the growth and remodelling of the vasculature. It has been demonstrated the role of VEGFB-VEGFR1 inducing the expansion of adipose vasculature and perfusion, increasing the resistance in diet-induced obesity[45].

Other important gene is the **LDLR** gene. This is the gene of the low-density lipoprotein receptor (LDL-R). This protein mediates the endocytosis of cholesterol-rich LDL and it is involved in leptin signalling[46]. Mutation in this receptor could cause familiar arteriosclerosis and hyperlipidaemia[47].

It has also been found evidence for association between the markers of the gene **TGFB1** (Transforming growth factor, β 1) with obesity-related phenotypes (BMI and sagittal abdominal diameter)[48].

The **PIK3CA** is the gene of the protein phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha. This protein is a subunit of the enzyme phosphatidylinositol 3-kinase (PI3K). Mice models without this subunit have shown reduced brown adipose activity. This produced a reduced energy expenditure, which promoted obesity and other metabolic consequences[49].

Table 15: significant SNVs which affect genes associated with phospholipid metabolic pathway. These SNVs are very common in obese samples but rare in controls except the highlighted.

Chr.	*Pos.*	*MAF*	*Ref.Alt*	*N_ob*	*TYPE*	*EU_MAF*	*p-value*	*Padj.fdr*	*GENES*	*MAF_TM*	*MAF_GMAD*
1	204140282	0.292	TCATC/T	12	INDEL	0.003	8.20E-07	1.86E-03	ETNK2	0.03	NA
2	98544045	0.344	GCA/G	16	INDEL	0	5.70E-11	7.63E-07	INPP4A	7.00E-04	0
3	**53189275**	0.3	T/G,*	15	SNP	0.002	7.45E-07	1.73E-03	PRKCD	**0.31**	**0.27**
3	179203465	0.219	G/GTAAA	16	INDEL	0	2.00E-07	5.47E-04	PIK3CA	8.00E-06	NA
4	53986526	0.444	[1]	9	INDEL	0	1.44E-08	6.74E-05	PDGFRA	0.009	NA
7	22961164	0.406	G/GCTCT	16	INDEL	0	5.70E-11	7.63E-07	FAM126A	NA	NA
8	131940480	0.367	A/AT,ATT	15	INDEL	0	3.60E-11	6.20E-07	EFR3A	2.00E-04	0
9	121183227	0.458	C/CA	12	INDEL	0	6.59E-10	5.71E-06	RAB14	2.00E-05	NA
9	133791944	0.364	[2]	11	INDEL	0	3.74E-08	1.37E-04	VAV2	0.002	NA
13	28368703	0.35	CTTT/CTT,C	10	INDEL	0	2.39E-08	9.24E-05	FLT1	8.00E-16	NA
16	68252675	0.308	GA/G	13	INDEL	0	8.00E-08	2.53E-04	PLA2G15	2.00E-05	NA
17	8881583	0.357	C/CAGTA	14	INDEL	0.001	9.73E-09	5.01E-05	PIK3R5	NA	0
19	3648502	0.25	[3]	14	INDEL	0	1.11E-07	3.42E-04	PIP5K1C	0.001	NA
19	11119974	0.308	GAAAC/G	13	INDEL	0	8.00E-08	2.53E-04	LDLR	NA	NA
19	41348575	0.269	TTTTA/T	13	INDEL	0	5.43E-06	9.53E-03	TGFB1	NA	NA
20	45419425	0.469	C/CG	16	INDEL	0	1.12E-14	9.82E-10	PIGT	NA	2.00E-04

[1] CTTTTATTTTATTTTATTTTATTTTA/C,CTTTTATTTTATTTTA,
[2] ACTGGGTGGGGTGTGTGTGCATGTGAGCGGGCTGTGCTGGGTGGGGGGTGTGTGACTGTGTGTGAATGAGCTGTG/A
[3] AGGCGCCCACCTGTGGGGCTGCAGACCCG/A

5.1.2. Genes related with platelet activation

Another biological process where the SNV-related genes are involved is the platelet activation. After GO enrichment, this biological process was the most significative (p-value = 1.78E-05). From the total amount of SNV-related genes, 11 of them are involved in platelet activation (ARRB1, GNAS, HRG, MYH9, PDGFRA, PIK3CA, PRKCD, PRKCG, PRKG1, VAV2, PIK3R5) of which 6 of them are new ones (Table 16) comparing with the last biological process discussed. Regarding the relationship between platelet activation and obesity, it has been found that the increased in-vivo parameter of platelet activation (mean platelet volume [MVP]) correlates positively with the BMI[50]. In addition, the biological and functional importance of some of these genes in relation to obesity have already been reported.

The importance of the gen **ARRB1,** which code for the β-arrestin-1 protein, has been dilucidated in knock-out mice. Knock-out of the gene encoding β-arrestin-1 caused increased fat mass accumulation and decreased whole-body insulin sensitivity in mice fed a high-fat diet. It was observed disrupted food intake and energy expenditure and increased macrophage infiltration in white adipose tissue. At molecular level, β-arrestin-1 deficiency affected the expression of many lipid metabolic genes and inflammatory genes in adipose tissue[51].

The gene **GNAS** has been related with early-onset obesity. GNAS is a complex imprinted locus with multiple oppositely imprinted gene products, including the G protein α-subunit Gsα which is expressed primarily from the maternal allele in some tissues and the Gsα isoform XLαs which is expressed only from the paternal allele. Mouse studies show that Gsα mutations lead to obesity due to Gsα imprinting in the central nervous system. This produce a specific defect in the ability of central melanocortins to stimulate sympathetic nervous system activity and energy expenditure[52,53].

5.1.3. Genes related with focal adhesion

From the KEGG enrichment, one the most significant pathway where the SNV-related genes are involved is the focal adhesion (p-value=5.026766e-05). 12 SNV-related genes are involved in this pathway (FLT1, FLT4, TNC, ITGA1, ITGA9, LAMA3, PDGFRA, PIK3CA, PRKCG, VAV2, PIP5K1C, PIK3R5) of which 5 of them are new ones (Table 17) comparing with the biological process before explained. The focal adhesion process is comprised of large molecular complexes which mediate signals modulating cell attachment, migration, proliferation, differentiation and gene expression[54]. This variety of functions make this pathway very important in several cell process and the impaired working of any protein that form the complexes can produce a wide range of diseases[54]. From the genes involved in this pathway, only TNC has been associated biologically and functionally with obesity.

The gene **TNC** encode the glycoprotein Tenascin C. This glycoprotein belongs to the damage-associated molecular patterns family and it is reported in the etiopathology of obesity via visceral adipose tissue inflammation representing a link with extracellular matrix remodelling[55,56].

5.1.4. Common SNVs in obese samples but rare in controls samples.

With the information about the MAF from samples and control, together with the information about the number of obese samples that presented the allele, it was determined those SNV that were very frequents in obese samples (MAF $\geq$ 0.45, N_ob $\geq$ 14) and very weird controls (MAF $\leq$ 0.01) (Table 9). In this section, it is explained the biological and functional importance of the genes, where these variants are located, in relation to obesity.

A high throughput small interfering RNA screen with the human primary adipocytes revealed that knockdown of human axonemal dyneins (DNAH) (such as **DNHA8** and **DNAH11)** decreased lipid accumulation[57]. This finding demonstrated that DNHAs are involved in fatty tissue biology and the impaired functioning of some of them can cause obesity[57].

The endothelial PAS domain protein 1 (**EPAS1**) is a transcription factor performantly expressed in endothelial cells and plays a role in the maintenance of reactive oxygen species. This protein has been related with adipogenesis because several growth factors, including insulin and insulin-like growth factor, induce and activate EPAS1 activity. In addition, EPAS1 is highly expressed in white adipose tissue and , more important if possible, the overexpression of EPAS1 in adipocytes has been significantly associated with the lipid accumulation[58].

5.1.5. Power of the findings and new genes associated with obesity

The fact that: [1] All molecular function and metabiotic pathways before explained are direct or indirectly related with obesity, [2] Some of the genes involved to those pathways have been correlated biologically and functionally with obesity and [3] All SNV-related genes involved in the molecular functions and metabolic pathways before explained had been associated with obesity according the GWAS database[35], showing how powerful these findings are. In addition, the results of this study suggest between 122 and 150 **new genes associated with obesity** and their respective variants (Table S3 and S4 in repository).

Table16: significant SNVs which affect genes associated with platelet activation. These SNVs are very common in obese samples but rare in controls except the highlighted.

Chr.	*Pos*	*MAF*	*N_ob*	*Ref.Alt*	*EUR_AF*	*pvalue*	*Padj.fdr*	*GENES*	*MAF_TM*	*MAF_GMAD*
11	75276977	0.385	13	G/GTCCCC	0	1.06E-09	7.46E-06	ARRB1	1.00E-04	NA
20	58899665	0.4	10	GCA/G	0.001	1.19E-07	3.55E-04	GNAS	0.001	NA
3	**186675306**	0.455	11	T/A	0.003	8.01E-09	4.20E-05	HRG	**0.39**	NA
3	**186675308**	0.3	10	T/A	0.003	2.27E-05	3.54E-02	HRG	**0.44**	NA
22	36326793	0.318	11	CG/C	0	3.16E-06	5.93E-03	MYH9	8.00E-05	NA
22	**36348845**	0.292	12	C/T	0.001	2.76E-07	7.36E-04	MYH9	**0.5**	NA
19	53892752	0.318	11	[1]	0.003	3.11E-05	4.44E-02	PRKCG	0.006	NA
19	53892758	0.292	12	A/G	0.002	2.76E-07	7.36E-04	PRKCG	0.05	NA
10	51320450	0.357	14	AG/A	0	1.64E-09	1.09E-05	PRKG1	8.00E-06	NA

[1] GCACA/G,GCACACA,GCA

Table 17: Significant SNV associated with focal adhesion genes.

Chr.	*Pos.*	*MAF*	*N_ob*	*Ref.Alt*	*EUR_AF*	*pvalue*	*Padj.fdr*	*GENES*	*MAF_TM*	*MAF_GMAD*
5	180621986	0.308	13	G/GCCTC	0	8.00E-08	2.53E-04	FLT4	1.00E-16	NA
9	115020557	0.417	12	T/TA	0.008	1.59E-07	4.64E-04	TNC	0.005	0.03
5	52952204	0.4	10	G/GA	0.008	2.91E-06	5.60E-03	ITGA1	8.00E-04	NA
3	37519086	0.273	11	CT/C,CTT	0	3.16E-06	5.93E-03	ITGA9	NA	NA
18	23813265	0.5	10	A/AT	0	2.13E-10	2.45E-06	LAMA3	2.00E-05	NA

5.2. CVN analysis discussion.

In this study 74 overlapping CNV regions have been identified. These regions are in a significantly different number of individuals (p-value ≤ 0.05) comparing case and controls samples. Of these 74 CNV regions, 72 are located in the chromosome 12. Some loci in chromosome 12 have been reported as associated with obesity related traits such as BMI, waist circumference, waist hip ratio (WHR) and weight[59].

5.2.1. Locus 12q13

The locus 12q13 have been associated with obesity because some important genes initiator and inhibitor of apoptosis such as FAIM2[60] or TMBIM6[61]are located there. These genes are related with obesity because of the role of leptin inducing adipose apoptosis as part of the actions aimed at controlling weight[15,62].

In the results, a total of 10 significant CNV regions are located in 12q13 locus. In the Figure 6 are exposed some of these CNV regions located in that locus.

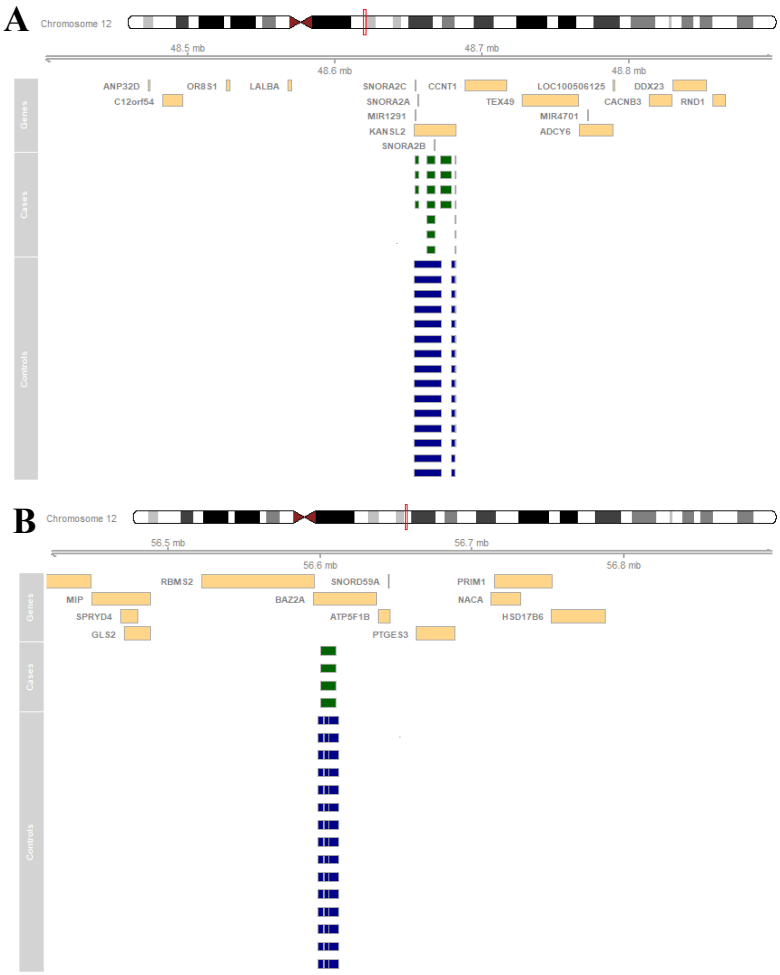

Figure 6: genome visualization of the regions chr12:48400000-48900000 **[A]** chr12:56420367-56900000 **[B]**. The frequency of CNVs regions among Cases (obese samples) represented as green boxes and Controls represented as blue boxes are shown

5.2.2. Locus 12q23-24

It has been found several evidences for linkage of obesity related phenotypes to markers in chromosome regions 12q23-24 with high level of significance[63]. In this region are located several genes such as IGF-I, SCARB1, ACACB, PMCH, C12orf51 and PTPN which are reported as associated with obesity[63–65].

In the results there are 16 different CNV regions which are located in 12q23-24 loci. In the figure 7 are exposed some of these regions located in 12q23-24 loci.

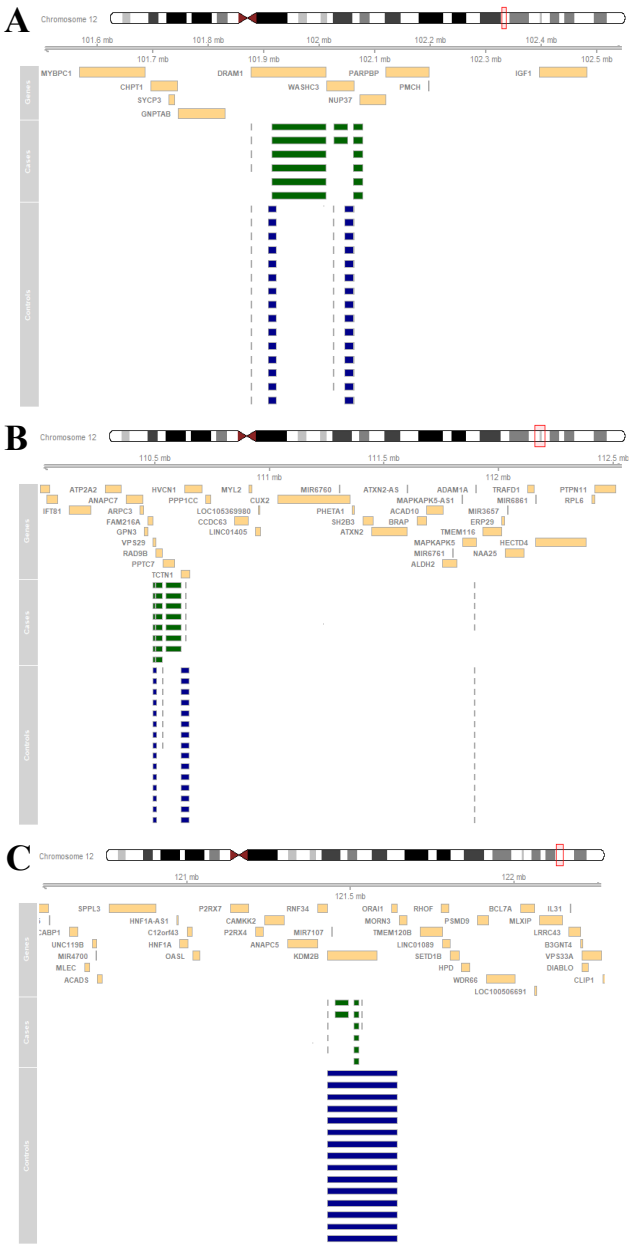

Figure 7: genome visualization of the regions chr12: chr12:101500000-102550000 **[A]** chr12:110000000-112550000 **[B]** and chr12:120550000-122275309 **[C]**. The frequency of CNV regions among Cases (obese samples) represented as green boxes and Controls represented as blue boxes are shown

From the CNV regions before exposed some important information can be inferred:

- The presence of ta CNV region is indicative of the presence of larger CNV which overlap with that region. So, it can be concluded that a lot of obese samples do not have the CNVs what controls samples do have. This could cause an abnormal expression of some obesity related genes. A gene expression analysis would be advisable in order to determine if there are any difference in gene expression between cases and controls.
- From the 74 significant CNV region, only 26 are located in loci which have already been related with obesity in previous studies. Hence, the other 47 CNV regions located in the loci 12p11, 12q15, 12q21 and 12q22 could be the evidences of relationship between these loci and obesity or, conversely, the data did not have the proper quality to perform the CNV procedure applied. In any case, new studies have to be performed over a larger number of samples in order to confirm or deny the findings.

6. Conclusion

Considering the objectives of this study, after performing al procedures and analysing the results, the following conclusions are extracted.

- From SNV detection procedure, 704 SNV which affected 479 different genes were obtained statistically correlated with obesity (p-value FDR adjusted ≤ 0.05). These variants have a very good quality because the majority of these genes had already correlated with obesity in previous GWAS studies.

- 329 of 479 genes have already been correlated with obesity. This suggests the discovery of 150 genes with their respective variants associated with obesity in Spanish population.

- From CNV analysis, a total of 73 CNV regions were found which were very frequent in controls but not in samples. 71 of them were located in the loci 12p11, 12q13, 12q15, 12q21, 12q22, 12q23 and 12q24.

- Only the loci 12q13, 12q23 and 12q24, where 26 CNV regions were located, had been correlated with obesity suggesting, on one hand, the possibility that new genomic regions correlated with obesity and, in the other hand, it could be possible that some bias into the data have produced wrong results.

In any case for both SNV detection and CNV analysis, extra studies have to be carried out to confirm or to reject the findings in this work. In addition, it should be taken into account that the procedures were performed with few numbers of samples (16 for SNV detection and 15 for CNV analysis). A larger number of samples would be preferable in subsequent analyses.

7. 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? *Diabetes*. 2003;52(2):232-238. doi:10.2337/diabetes.52.2.232

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 HH, Neale MC, Eaves LJ. Genetic and environmental factors in relative body weight and human adiposity. *Behav Genet*. 1997;27(4):325-351. http://www.ncbi.nlm.nih.gov/pubmed/9519560.

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. *Nat Genet*. 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. *Proc Natl Acad Sci*. 2003;100(16):9440-9445. doi:10.1073/pnas.1530509100

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. Li MJ, Wang P, Liu X, et al. GWASdb: A database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res*. 2012;40(D1):1047-1054. doi:10.1093/nar/gkr1182

36. Li MJ, Wang P, Liu X, et al. Gene Set - obesity. http://amp.pharm.mssm.edu/Harmonizome/gene_set/obesity/GWASdb+SNP-Disease+Associations. Accessed June 25, 2019.

37. 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

38. 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

39. 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.

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

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

42. 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

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

44. 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

45. Robciuc MR, Kivelä R, Williams IM, et al. VEGFB/VEGFR1-Induced Expansion of Adipose Vasculature Counteracts Obesity and Related Metabolic Complications. *Cell Metab*. 2016;23(4):712-724. doi:10.1016/j.cmet.2016.03.004

46. Ngai YF, Quong WL, Glier MB, et al. Ldlr-/- mice display decreased susceptibility to western-type diet-induced obesity due to increased thermogenesis. *Endocrinology*. 2010;151(11):5226-5236. doi:10.1210/en.2010-0496

47. Coenen KR, Gruen ML, Hasty AH. Obesity causes very low density lipoprotein clearance defects in low-density lipoprotein receptor-deficient mice. *J Nutr Biochem*. 2007;18(11):727-735. doi:10.1016/j.jnutbio.2006.12.010

48. Pérusse L, Rankinen T, Zuberi A, et al. The Human Obesity Gene Map: The 2004 Update. *Obes Res*. 2005;13(3):381-490. doi:10.1038/oby.2005.50

49. Nelson VLB, Ballou LM, Lin RZ. Energy balancing by fat Pik3ca. *Adipocyte*. 2015;4(1):70-74. doi:10.4161/21623945.2014.955397

50. Santilli F, Vazzana N, Liani R, Guagnano MT, Davì G. Platelet activation in obesity and metabolic syndrome. *Obes Rev*. 2012;13(1):27-42. doi:10.1111/j.1467-789X.2011.00930.x

51. Zhuang LN, Hu WX, Zhang ML, et al. B-Arrestin-1 Protein Represses Diet-Induced Obesity. *J Biol Chem*. 2011;286(32):28396-28402. doi:10.1074/jbc.M111.223206

52. Grüters-Kieslich A, Reyes M, Sharma A, et al. Early-onset obesity: Unrecognized first evidence for GNAS mutations and methylation changes. *J Clin Endocrinol Metab*. 2017;102(8):2670-2677. doi:10.1210/jc.2017-00395

53. Weinstein LS, Xie T, Qasem A, Wang J, Chen M. The role of GNAS and other imprinted genes in the development of obesity. *Int J Obes*. 2010;34(1):6-17. doi:10.1038/ijo.2009.222

54. Lijnen HR. Angiogenesis and obesity. *Cardiovasc Res*. 2008;78(2):286-293. doi:10.1093/cvr/cvm007

55. Catalán V, Gómez-Ambrosi J, Rodríguez A, et al. Increased tenascin C and toll-like receptor 4 levels in visceral adipose tissue as a link between inflammation and extracellular matrix remodeling in obesity. *J Clin Endocrinol Metab*. 2012;97(10):1880-1889. doi:10.1210/jc.2012-1670

56. Catalán V, Gómez-Ambrosi J, Rodríguez A, et al. Increased obesity-associated circulating levels of the extracellular matrix proteins osteopontin, chitinase-3 like-1 and tenascin C are associated with colon cancer. *PLoS One*. 2016;11(9):1-15. doi:10.1371/journal.pone.0162189

57. Söhle J, Machuy N, Smailbegovic E, et al. Identification of new genes involved in human adipogenesis and fat storage. *PLoS One*. 2012;7(2). doi:10.1371/journal.pone.0031193

58. Shimba S, Wada T, Hara S, Tezuka M. EPAS1 promotes adipose differentiation in 3T3-L1 cells. *J Biol Chem*. 2004;279(39):40946-40953. doi:10.1074/jbc.M400840200

59. Alonso R, Farías M, Alvarez V, Cuevas A. The Genetics of Obesity. *Transl Cardiometabolic Genomic Med*. 2015;(October):161-177. doi:10.1016/B978-0-12-799961-6.00007-X

60. Li C, Qiu X, Yang N, et al. Common rs7138803 variant of FAIM2 and obesity in Han Chinese. *BMC Cardiovasc Disord*. 2013;13(1):56. doi:10.1186/1471-2261-13-56

61. Bailly-Maitre B, Belgardt BF, Jordan SD, et al. Hepatic Bax inhibitor-1 inhibits IRE1α and protects from obesity-associated insulin resistance and glucose intolerance. *J Biol Chem*. 2010;285(9):6198-6207. doi:10.1074/jbc.M109.056648

62. Gullicksen PS, Della-Fera MA, Baile CA. Leptin-induced adipose apoptosis: Implications for body weight regulation. *Apoptosis*. 2003;8(4):327-335. doi:10.1023/A:1024112716024

63. Li WD, Dong C, Li D, Zhao H, Price RA. An Obesity-Related Locus in Chromosome Region 12q23-24. *Diabetes*. 2004;53(3):812-820. doi:10.2337/diabetes.53.3.812

64. Pérusse L, Rice T, Chagnon YC, et al. A genome-wide scan for abdominal fat assessed by computed tomography in the Québec Family Study. *Diabetes*. 2001;50(3):614-621. doi:10.2337/diabetes.50.3.614

65. Wilson SG, Adam G, Langdown M, et al. Linkage and potential association of obesity-related phenotypes with two genes on chromosome 12q24 in a female dizygous twin cohort. *Eur J Hum Genet*. 2006;14(3):340-348. doi:10.1038/sj.ejhg.5201551