**UNIVERSIDAD AUTÓNOMA DE MADRID**

**ESCUELA POLITÉCNICA SUPERIOR**



**TRABAJO FIN DE MÁSTER**

# EVALUACIÓN DE MÉTODOS ESTADÍSTICOS EN LA ASOCIACIÓN DE VARIANTES RARAS Y RIESGO DE CÁNCER DE VEJIGA.

**Máster Universitario en Bioinformática y Biología Computacional**

**Autor: Fernández del Pozo, Alba**

**Tutor: López de Maturana López de Lacalle, Evangelina**
**Departamento: Epidemiología Genética y Molecular. Centro Nacional de Investigaciones Oncológicas (CNIO)**

# EVALUACIÓN DE MÉTODOS ESTADÍSTICOS EN LA ASOCIACIÓN DE VARIANTES RARAS Y RIESGO DE CÁNCER DE VEJIGA.

Autor: Alba Fernández del Pozo
Tutor: Evangelina López de Maturana López de Lacalle
Ponente: Luis del Peso Ovalle

Escuela Politécnica Superior
Universidad Autónoma de
Madrid MES 20

## Resumen

**Introducción**. El cáncer de vejiga es una enfermedad compleja. Los estudios de asociación de todo el genoma (GWAS), han identificado variantes comunes asociadas con el riesgo de desarrollar cáncer de vejiga. Sin embargo, la implicación de las variantes raras aún no se ha explorado. Aquí exploramos la contribución de los SNP de codificación rara al componente genético del desarrollo de cáncer de vejiga.

**Material y Métodos**. Toda la secuenciación del exoma se realizó en el ADN germinal de 104 individuos (68 casos/36 controles) que participaban dentro del estudio EPICURO para cáncer de vejiga. Los sujetos fueron seleccionados siguiendo un diseño de fenotipo extremo con el fin de potenciar el componente genético. Solo las variantes con una frecuencia de alelo menor <0.01 y buena calidad de imputación (info > 0.3) se analizaron utilizando tres métodos de colapso: Burden test, SKAT y SKAT-O, en función de los genes. Priorizamos los genes identificados por los tres métodos después de realizar la corrección por pruebas múltiples.

**Resultados**. Un número total de 93,867 variantes raras anotadas en ~ 14,700 genes finalmente se incluyeron en el análisis. Burden test identificó 184 genes, mientras que SKAT y SKAT-O identificaron 169 y 197 genes, respectivamente, con 119 genes identificados por los tres métodos. Curiosamente, dos de ellos, *LIG1* y *ERCC1*, se asociaron previamente con neoplasias de vejiga. Además de ellos, se identificaron nuevos posibles genes de susceptibilidad. Después de anotarlos en las vías de KEGG, los más relevantes fueron: "Reparación de escisión de nucleótidos", que tiene otros genes de susceptibilidad a cáncer de vejiga como *XPC* y *XPD*, también se relacionó con cáncer de vejiga en un análisis de grupo anterior, "Regulación del citoesqueleto de actina", que tiene *FGFR3*, otro gen de susceptibilidad para cáncer de vejiga, y "Estrecha unión", que tiene 3 posibles genes de susceptibilidad encontrados en nuestro análisis.

**Conclusiones**. La gran cantidad de genes seleccionados por los tres métodos sugiere que las variantes raras heredadas en muchos genes contribuyen al riesgo de desarrollar cáncer de vejiga. Hasta donde sabemos, este es el primer estudio que investiga el papel de las variantes raras en la susceptibilidad genética al cáncer de vejiga.



Palabras Clave: Cáncer, Vejiga, MAF, Rare, Burden, SKAT, SKAT-O, SeqMeta.

## *Abstract Key*

**Background.** Bladder cancer (BC) is a complex disease. Genome-wide association studies have identified common variants associated with BC risk. However, the implication of rare variants has not been explored yet. Here we explored the contribution of rare coding SNPs to the genetic component of BC development**.**

**Methods.** Whole exome sequencing was conducted in germline DNA from 104 individuals (36 controls/68 cases) participating in the Spanish Bladder Cancer/EPICURO study. Subjects were selected following an extreme phenotype design in order to potentiate the genetic component. Only those variants with a minor allele frequency <0.01 and good quality of imputation (info>0.3) were analyzed using three collapsing methods (Burden test, SKAT and SKAT-O) on a gene-basis. We prioritized the genes identified by the three methods, after multiple testing correction.

**Results.** A total number of 93,867 rare variants annotated on ~14,700 genes were finally included in the analysis. The Burden test identified 184 genes, while SKAT and SKAT-O identified 169 and 197 genes, respectively, with 119 genes being identified by the three methods. Interestingly, two of them, *LIG1* and *ERCC1,* were previously associated with bladder neoplasms. In addition to them, novel possible susceptibility genes were identified. After annotated them in KEGG pathways, the most relevant ones were: “Nucleotide excision repair”, which has other BC susceptibility genes as *XPC* and *XPD*, also reported to be associated with BC in a previous pool analysis, “Regulation of actin cytoskeleton”, which has *FGFR3*, another susceptibility gene for BC, and “Tight junction”, which has 3 possible susceptibility genes found in our analysis.

**Conclusions.** The large number of genes selected by the three methods suggests that rare inherited coding variants across many genes contribute to BC risk. To our knowledge, this is the first study investigating the role of rare variants on the genetic susceptibility to BC.



*Words:* Cancer, Bladder, MAF, Rare, Burden, SKAT, SKAT-O, SeqMeta.

# Agradecimientos

**A Evangelina,**

por todo su tiempo y dedicación, su paciencia y su enorme disposición a ayudarme en cualquier momento y ante cualquier problema, gracias por todo lo que me has enseñado y lo que me llevo aprendido, y por darme esa soltura para desenvolverme en un mundo nuevo para mí.

**A Núria,**

por su confianza en mi desde el primer día, sus propuestas y consejos de mejora, y por dejarme participar de este maravilloso proyecto que me ha abierto las puertas a un campo lleno de nuevas oportunidades.

**A Marta,**

por su amabilidad y por el trato que me dio desde el primer momento, gracias por esa perspectiva estadística y por enseñarme a mirar las cosas desde otro punto de vista.

**A Lola,**

por su enorme generosidad, por regalarme momentos y enseñanzas, comandos, trucos y atajos para llegar al final del camino, por esa sonrisa tan necesaria en algunos días y por esa conversación siempre tan interesante y entretenida.

Y a todo **el grupo de Epidemiología Genética y Molecular del CNIO** por tratarme como una más y hacer que desde el primer día me sienta como en casa y que este proyecto haya formado parte de una etapa muy feliz tanto en mi vida profesional como personal.

Gracias a los PIs, investigadores, monitores y pacientes del estudio SBC/EPICURO. Mencionar a ESGI Project que financió la secuenciación, y al FIS y NCI que financiaron el estudio SBC/EPICURO.

GRACIAS

# Index

## Tables and Figures Index

# 1. INTRODUCTION

Bladder cancer (BC), ranks first among urinary carcinomas, representing 50% of the cases, ahead of other conditions such as kidney and renal pelvis cancer or that of ureters and other urinary organs [Siegel et al., 2017].

Figure 1 shows the worldwide incidence of bladder cancer. It is estimated that there are about 429,000 new cases/year worldwide, which rank the bladder carcinoma as the ninth most common cancer, for both sex combined, ahead of non-Hodgkin lymphoma, leukemia, pancreatic cancer or kidney cancer [Ferlay et al., 2015]. In the European Union (EU), 124,000 people are diagnosed of BC each year and, by 2030, the annual incidence of this cancer is projected to increase to 219,000 two-fifths of this as a consequence of the ageing of the European population [GLOBOCAN 2012].

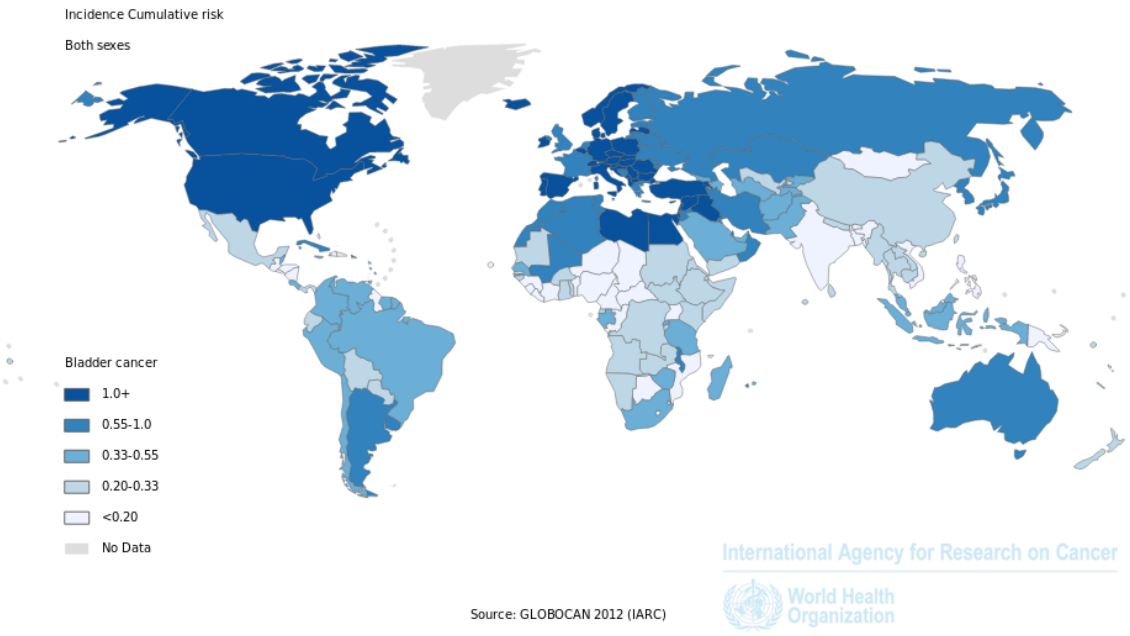


**Figure 1. Worldwide incidence of bladder cancer**

Within the new cases per year, 330.000 (76.9%) correspond to men, with a male:female ratio of 3.5:1, worldwide [Ferlay et al., 2015].

In Spain, bladder cancer ranks fifth among the most frequent tumors diagnosed in 2015, with 21.093 cases, and a male-female ratio higher than that found worldwide, and which stands at 4.8:1 [SEOM, 2017].

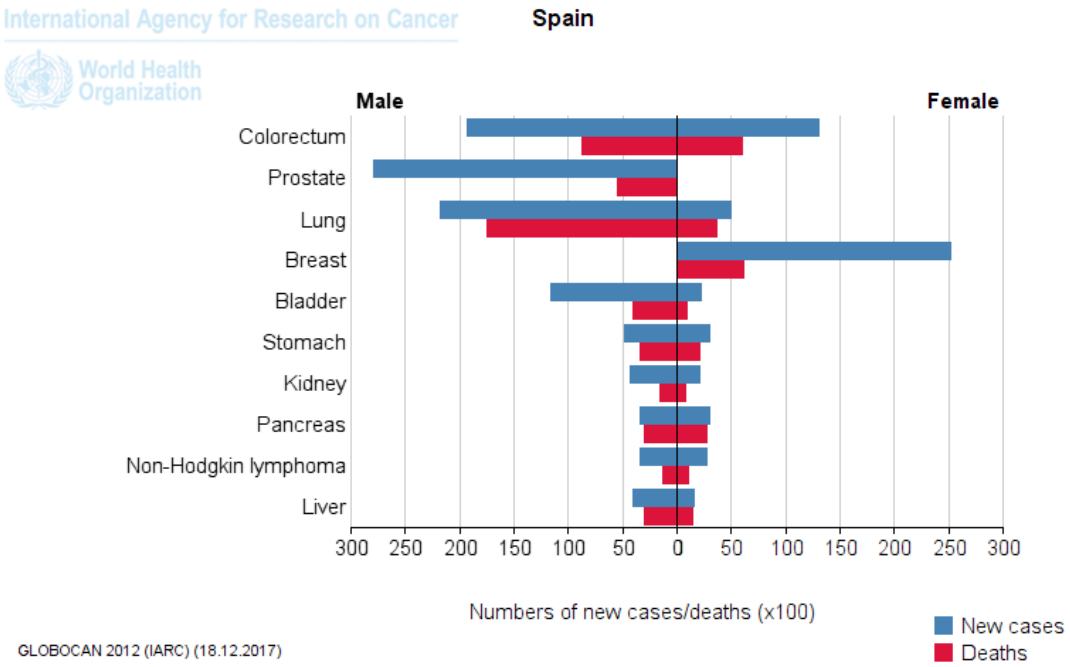


**Figure 2. Number of new cases/deaths by sex, in different cancers in Spain.**

Bladder cancer poses an economic burden to the EU, costing €4.9 billion in 2012 [Leal et al., 2016]. The five most populous countries including Spain, account for 73% of all costs, resulting in €3.6 billion. In particular, the BC health care cost for the EU health care systems was estimated as €2.9 billion in 2012, representing 59% of the total economic burden. In Spain, the annual health care costs of BC were equivalent to €61 per every 10 citizens, a higher cost than the one estimated for the whole EU (€57 per every 10 EU citizens) [Leal et al., 2016]. Another figure showing the importance of BC as an economic burden for the Spanish public health care system is that it represents >4% of all total cancer costs, whereas its counterpart in other European countries account for 2% of the total costs.

Planning urologic care systems requires a good knowledge of BC epidemiology. In this regard, BC is a paradigm of complex disease, and results from the interplay between both genetic and environmental effects.

Many studies have determined that smoking is the main risk factor for BC [Samanic et al., 2006] [Guillaume et al., 2014], with an attributable fraction of approximately 50% of the cases [Jankovic et al., 2007], since there is a direct relationship between the development of UBC and the content of aromatic amines such as B-naphthylamine and the polycyclic aromatic hydrocarbons of tobacco [Burguer et al., 2013].

Occupational exposure to these aromatic amines and polycyclic aromatic hydrocarbons are another of the most important risk factors for BC [Burguer et al., 2013], with up to 20% of UBCs being associated with exposure to these components in industrial areas. Furthermore, there are other less determinant factors such as diet and environmental pollution.

Regarding genetic factors, it has been found that the risk of BC is two-fold higher in first-degree relatives of patients diagnosed with BC [Burguer et al., 2013]. Although familial aggregation of BC has been described, no high-penetrance allele/gene has been identified so far explaining these familial clusters. Genome-wide association studies (GWAS) have shed some light on the deciphering of the genetic susceptibility of BC. So far, 24 loci distributed in 19 regions of 14 chromosomes (see Figure 3) have been identified according to GWAS catalog [Welter et al., 2014] [Figueroa et al., 2016] explaining ~12% of the familial risk. This fact along with the low incidence of BC and the nonexistent cost-effective measures, make impracticable to establish a screening strategy for BC in asymptomatic adults, based only on genetics information [Moyer, VA. 2006]. However, it is necessary to identify those risk groups and prioritize the allocation of public funds for research [Leal et al., 2015].

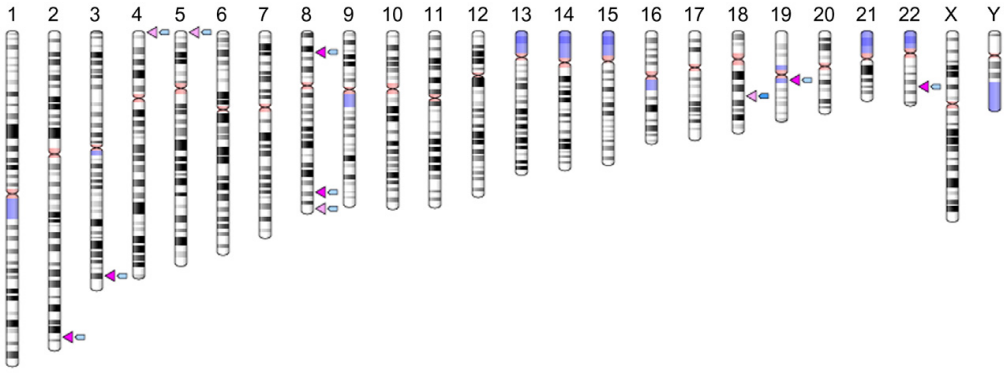


**Figure 3. Ideogram localizing the GWAS identified common genetic susceptibility variants in the chromosome bands. From the Phenotype-Genotype Integrator (PheGenI) after searching for urinary bladder neoplasm.** [López de Maturana et al., 2017]

The most relevant genetic factors in terms of conferring risk are the slow acetylator N-acetyltransferase 2 (*NAT2*) and the glutathione S-transferase mu 1 (*GSTM1*) null genotypes, well established risk factors for BC [García-Closas et al., 2005]. These variants confer individual susceptibility to exogenous carcinogens, mainly those present in tobacco, since both enzymes are involved in the detoxification of such carcinogenss [Burguer et al., 2013]. In a combined analysis, the association between BC risk and *GSTM1* deletion was stronger in never smokers (OR=1.75), and progressively weaker in former (OR=1.55) and current smokers (OR=1.25); on the other hand,

the association between *NAT2* and BC risk was limited to cigarette smokers (OR=1.24) [Rothman et al., 2010].

Genome-wide association studies have been proven to be useful tools to identify genetic associations between common variants and disease [Manolio et al., 2009]. Regarding BC, all the genetic associations found to date through candidate gene approaches or GWAS, correspond to low penetrant variants that appear frequently in the population (e.g., common variants, with a frequency of the minor allele larger than 5%) (See Figure 4). However, little is known about the contribution of rare variants to BC genetic susceptibility. Rare variants (MAF <0.5%) are known to play an important role in human diseases and their study could offer an explanation about the variability and risk of suffering certain diseases, as BC [Seunggeung et al., 2014]. The rare allele model assumes that there are many rare alleles of large effect, so the disease would be due to rare variants of high penetrance [Gibson, 2015]. In the case of rare variants, GWAS are not useful to detect associations with rare variants, because common variants do not capture the genetic variation due to rare variants [Gibson, 2010].

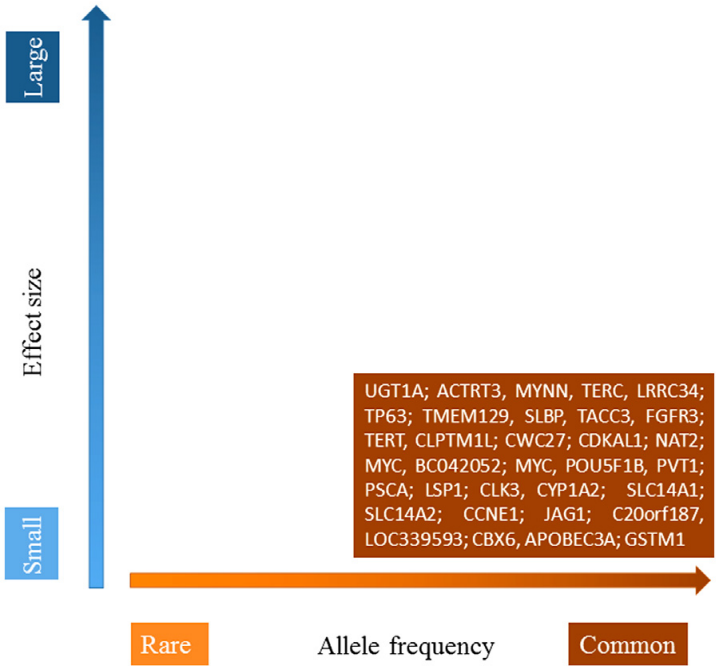


**Figure 4. Genetic susceptibility genes associated with urothelial bladder cancer in the GWAS catalogue displayed according to their minor allele frequency (x-axis) and effect size (y-axis).** [López de Maturana et al., 2017]

Next generation sequencing (NGS) technologies are parallel-sequencing approaches that generate billions of short sequence reads at a modest cost. In recent years, NGS has been a revolution in the field of genomics, allowing genome-wide sequence data to be generated quickly and cost-effectively with high accuracy [Jiekun et al., 2015]. These short reads are aligned to a reference genome given the possibility to identify and genotype sites where sequenced individuals vary [Seunggeung et al., 2014]. The continuous decrease in the price of this technique, has made of

NGS a powerful and indispensable tool in the genetic association studies, by enabling a more complete assessment of the role of low-frequency and rare genetic variants in complex traits. However, the detection of rare variants in sequencing-based approaches is challenging. First, performing NGS in a large sample of individuals is costly. In order to minimize this limitation, various alternative strategies have been proposed: targeted sequencing, exome sequencing, low-depth WGS, and extreme-phenotype sampling.

The second challenge relates to the limited power of classical statistical tests to detect rare associations with rare variants. Mainly due to their low frequency, the proportion of the genetic variation explained by these low frequent variants is likely to be small, unless their effect is very large, impairing their detection with classical statistical approaches.

Because the number of rare variants is much greater than that of common ones, stricter levels of significance could be required, reducing statistical power [Seunggeung et al., 2014]. In order to overcome these limitations, novel statistical methods have been recently developed, that instead of testing each rare variant individually, they collapse them, and evaluate their joint effect on a gene or biologically relevant region basis, increasing the power when there is association with the disease. [Seunggeung et al., 2014]. The aggregation tests developed so far differ on the assumptions about underlying genetic model, and therefore, their power depends on the true genetic model, which in real data is likely to be unknown. According to their assumptions, aggregation methods can be categorized as: burden tests, adaptive burden tests, variance-component tests, combined burden and variance-component tests, and the exponential-combination (EC) test [Seunggeung et al., 2014].

Specifically, some of the most used methods for the analysis of rare variants are: Burden test [Li et al., 2008], Sequence Kernel Association Test (SKAT) [Wu et al., 2011], and Sequence Kernel Association test-Optimized (SKAT-O) [Wang, 2016].

Here we explored for the first time the contribution of rare coding SNPs to the genetic component of BC development using the current statistical methods for the analysis of rare variants, and a whole exome sequencing approach (WES).

## 2. OBJETIVES

The overall aim of this project is to explore and compare the performance of the current statistical methods for the analysis of rare variants in assessing the contribution of rare coding SNPs to the genetic susceptibility of BC using a WES-based approach.

The specific objectives are:

1) To identify the statistical methods currently available for the association analysis of rare variants with risk.

2) To implement these methods in real data (exome sequencing data in BC) to identify rare, putatively functional, protein-coding variants associated with BC.

# 3. MATERIAL AND METHODS

## 3. 1. Study population

Here we used the resources of the SBC/EPICURO study, a retrospective hospital-based case–control study conducted in 18 hospitals in five Spanish areas (Asturias, Barcelona metropolitan area, Vallès/Bages, Alicante and Tenerife) during the years 1998 to 2001. [García-Closas M et al., 2005].

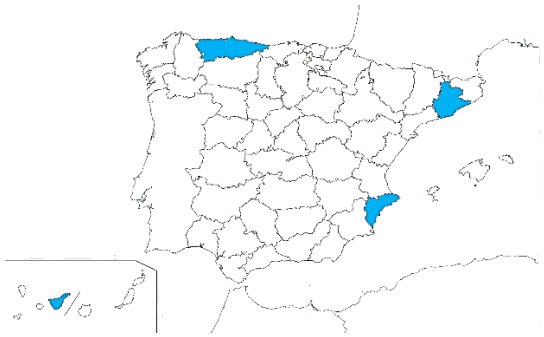

**Figure 5. Distribution areas of EPICURO study**

The inclusion criteria for cases were: age between 21-80 years old, current diagnosis of urinary bladder transition cell carcinoma, histologically confirmed, and classified as such by the 1998 system of the World Health Organization (WHO) and the International Society of Urological Pathology [Epstein et al., 1998].

The controls were patients admitted in the participating hospitals for other pathologies, whose diagnosis were not related to the known risk factors of BC. Controls were matched to a case for to age (in categories of 5 years), sex, ethnicity and region. The informed consent of the participants was obtained in accordance with the Institutional Review Board of the US National Cancer Institute and the Ethics Committees of each participating hospital.

Blood samples were requested for DNA extraction and information about risk factors (smoking habit, cancer history, family history of cancer and environmental exposure) , was collected trough personal interviews conducted by trained monitors. Subjects were also categorized according to their smoking habit in: never smokers; smoked less than 100 cigarettes in their lifetime; occasional smokers; smoked at least one cigarette per day for less than 6 months; former smokers if they had smoked regularly but stopped smoking more than one year before the study inclusion date, and current smokers, if they had smoked regularly within a year of the inclusion date [Samanic et al., 2016].

For the current project, those patients selected following an extreme phenotype design were used, in order to accentuate the genetic component in UBC, and assuming that rare variants were enriched among thoses cases. Given the binary nature of the phenotype data, subjects were selected according to their age at diagnosis/recruitment (≤50 years), first-degree relatives with cancer, and non-tobacco consumption. Ccontrols were >70 years old and heavy smokers (more than 21 cigarettes/ day) individuals without bladder and family cancer.This design was already used in a previous study, improving the classification performance of yet-to-be observed data using a multimarker model [López de Maturana et al., 2014]. Finally, data from 104 patients (68 cases and 36 controls) were used in the analysis.

Whole exome sequencing was conducted in germline DNA from these patients. WES provides coverage of about 95% of the exons, which contains 85% of mutations causing disease in Mendelian disorders throughout the genome [Rabbani B et al., 2014]. In addition, another advantage of the WES is that can be applied to moderate numbers of samples due to its lower price with respect to the complete genome.

## 3. 2. Whole exome sequencing workflow

Figure 6, shows the WES workflow used to generate the files, I used as starting point in this TFM (process 6 of the workflow). Briefly, DNA samples were sequenced using the Illumina TruSeq exome enrichment kit at the University of Uppsala (Sweden), within the framework of EXOMxPRED project. Reads were aligned to the human reference sequence (hg19) using the Burrows-Wheeler Aligner tool (BWA2) and the resulting binary alignment map (BAM) files were used to call single-nucleotide variants (SNVs) across all samples, i.e., multi-sample calling.

The VCF files generated were then processed in the Genetic and Molecular Epidemiology Group (GMEG), Spanish National Cancer Research Center (CNIO) using a pipeline developed by a previous Master student, Laura Leroi, to ensure the quality of the variant calling. The quality control procedure was applied at both sample and SNP level to ensure the quality of SNPs genotypes. Those with bad quality (e.g., SNPs with read depth < 10 and those which did not pass the tests of base sequencing quality, strand bias or tail distance bias) were considered as missing, and then imputed using a strategy of pre-phasing of the target dataset implemented in the framework IMPUTE2 plus SHAPEIT2 in the GMEG (CNIO). The imputation was performed combining the samples with WES information with the rest of individuals included in the SBC/EPICURO study that were genotyped with the 1M Illumina array [Rothman et al, 2010].

These imputed files were the initial files that were used in this TFM, from which we selected the autosomal chromosomes.

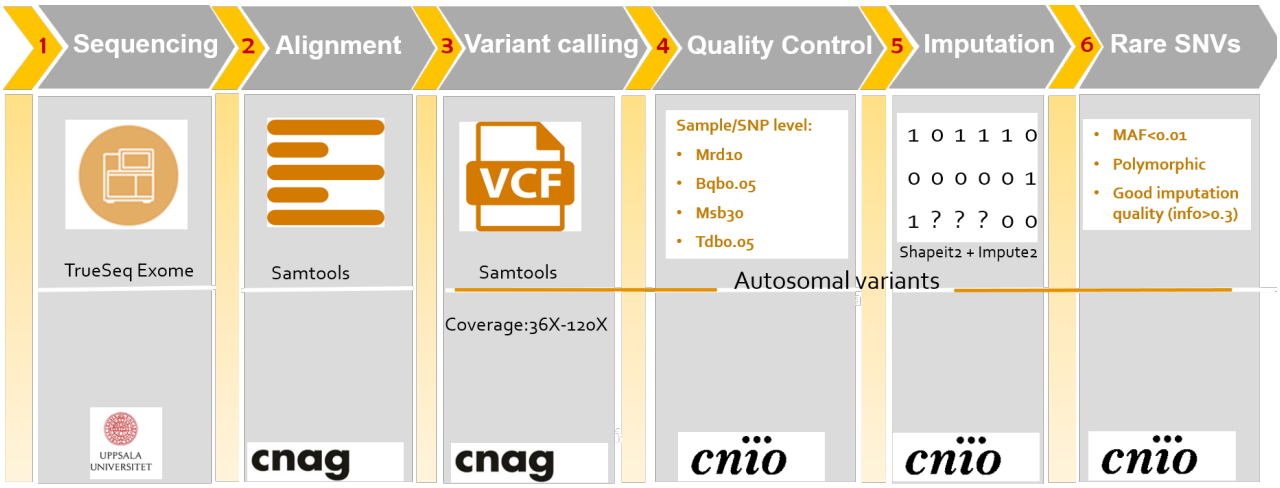


**Figure 6. Data generation and processing Workflow**

## 3. 3. Annotation

To perform the annotation of the SNPs we used the Genome Reference Consortium Human Build 37 (GRCh37), also known as hg19, available in Biomart, Ensemble (https://www.ensembl.org). The database *Ensembl Genes 91*, and the dataset *Human genes (GRCh37.13)* were selected and the output was saved in the file `ensembl_allGenes.bed`

The annotation was made by the script **'annotateVCF_ENSgenes.sh',** which is shown below**.**

First, I sorted the file `ensembl_allGenes.bed` by chromosome position and compressed the file. I used tabix tools to index and convert it into a VCF file. This VCF file was used as input of `vcf-annotate` (http://vcftools.sourceforge.net/perl_module.html) to perform the gene annotation.

```
#!/bin/bash
sort -k1,1V-k2,2n-V ensembl_allGenes.bed > sorted_ensembl_allGenes.bed
bgzip sorted_ensembl_allGenes.bed-c > sorted_ensembl_allGenes.bed.gz
tabix -s 1 -b 2 -e 3 -f sorted_ensembl_allGenes.bed.gz
for chr in $(seq 1 22); do
  awk 'BEGIN{FS="\t"; OFS="\t" {print $2,$3,$1,$4,$5,".",".","gene="$8,"."}'
EPICURO.${chr}/info.QC.only.SNP.EPICURO.${chr}.snpEff.p.SAL.SAL10_2.vcf.txt |
sed '1d' | awk 'BEGIN{print
"#CHROM\tPOS\tID\tREF\tALT\tQUAL\tFILTER\tINFO\tFORMAT"}; {print}; ' >
~/tfm_Alba/annotateVCF/chrom${chr}_vars.vcf;
done
for chr in $(seq 1 22); do
```

```
  cat /home/amfernandez/tfm_Alba/annotateVCF/chrom${chr}_vars.vcf | vcf-annotate
--annotations sorted_ensembl_allGenes.bed.gz --columns CHROM,FROM,TO,INFO/GENE -
description key=INFO,ID=GENE,Number=1, Type=String, Description='Gene Name'
> ./gene_annotated_output${chr}.vcf
done
```

**Script 'annotateVCF_ENSgenes.sh'**

## 3. 4. File Edition

The information needed to pursue the first objective of this TFM was stored in the following files:

A) `Info.QC.only.SNP.EPICURO.$.snpEff.p.SAL.SAL10_2.vcf` (where `$` corresponds to each autosomal chromosome). This is one of the output files obtained after applying the QC pipeline to the WES data from the individuals with extreme phenotypes (see the WES workflow in previous section for more details). It has the following format:

| | Name | Chr | Pos | Ref | Alt | Tri_all | TiTv_Type | Gene | Effect | Effect_lvl | Morphism | DP | Call_Rate | MAF | GMAF | Variant_Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | rs191348624 | 1 | 762061 | T | A | 0 | Tv | LINC00115 | EXON | MODIFIER | S | 3883 | 90.3846 | 0.0053 | 0.0023 | 3 |
| 3 | abc1var178 | 1 | 762109 | C | T | 0 | Ti | LINC00115 | EXON | MODIFIER | M | 10086 | 60.5769 | 0.0000 | NA | 3 |
| 4 | rs3115849 | 1 | 762273 | G | A | 0 | Ti | LINC00115 | EXON | MODIFIER | P | 5276 | 94.2308 | 0.8061 | 0.2770 | 1 |
| 5 | rs150580910 | 1 | 777318 | C | T | 0 | Ti | LINC01128 | DOWNSTREAM | MODIFIER | P | 5455 | 100.0000 | 0.0144 | 0.0119 | 2 |
| 6 | rs2980318 | 1 | 777361 | T | C | 0 | Ti | LINC01128 | DOWNSTREAM | MODIFIER | P | 6296 | 98.0769 | 0.0147 | 0.0371 | 2 |
| 7 | rs142849724 | 1 | 783071 | C | T | 0 | Ti | LINC01128 | EXON | MODIFIER | P | 4152 | 91.3462 | 0.0105 | 0.0101 | 2 |
| 8 | abc1var183 | 1 | 792467 | G | A | 0 | Ti | LINC01128 | DOWNSTREAM | MODIFIER | S | 3071 | 79.8077 | 0.0060 | NA | 3 |
| 9 | rs2905036 | 1 | 792480 | C | T | 0 | Ti | LINC01128 | DOWNSTREAM | MODIFIER | P | 3213 | 82.6923 | 0.9942 | 0.0165 | 1 |
| 10 | rs41285790 | 1 | 865628 | G | A | 0 | Ti | SAMD11 | NON_SYNONYMOUS_CODING | MODERATE | S | 4699 | 98.0769 | 0.0049 | 0.0032 | 3 |

**Figure 7. Info file from WES.**

Where,

- *Name:* Corresponds to the SNP id. For those SNPs without rs number, a unique id was assigned.
- *Chr:* Is the chromosome where each SNP is located.
- *Pos*: Corresponds to the SNP position in the chromosome.
- *Ref*: Is the reference allele.
- *Alt*: Corresponds to the alternative allele.
- *Tri_all*: To know whether if the variant has 2 different alternate base
- *TiTv_Type*: Indicates the type of the variation, where Ti corresponds to a transition and Tv to a tranversion.
- *Gene*: The name of the gene overlapping this position in HGNC gene symbol format, which was obtained after the annotation performed in CNAG.
- *Effect* : Effect of this variant from snpEff. See http://snpeff.sourceforge.net/
- *Effect_lvl:* Effects categorized by 'impact'. For more details, please see http://snpeff.sourceforge.net/

- *Morphism*: Type of polymorphism: singleton (S), monomorphic (M), and polymorphic (P).
- *DP*: This number indicates the total read depth for this position.
- *Call_Rate*: Percentage of genotype calls in the vcf file for this variant.
- *MAF*: Indicates the Minor Allele Frequency.
- *GMAF*: Is the Global Minor Allele Frequency provided by CNAG using 1000 Genomes database.
- *Variant_Type*: Rare, low frequency or common polymorphism.

B) `Info.$.chr.inf.` This is one of the output files obtained after the imputation procedure (see previous section for more details). It contains the following information: name of the SNP (*snp_id*), position (*rs_id*), major allele (*a0*), minor allele (*a1*), expected frequency of the alternative allele (exp_freq_a1), imputation quality measurement 1 (*info*), imputation quality measurement 2 (*certainty*), (*type;* 3 types: 0, SNP only in the reference, 2: SNP both in the reference and in our study*)*, 3: SNPs appearing only in our study), imputation quality measurement 3 (info_*type0*), imputation quality measurement 4 (*concord_type0*), imputation quality measurement 4 (*r2_type0*).

C) `imputed_dosages.$.chr.dsg.` Output file from the imputation procedure. It contains the SNP genotypes of all the genotyped individuals in SBC/EPICURO study (1127 cases/1090 controls), including the individuals with extreme phenotypes. It contains the genotypes resulting from the imputation (each column corresponded to an individual) from the 6th column; corresponding the first 5 fields to the [–, First name, Position, Most likely nucleotide, Alternative]

```
--- 1:10177:A:AC 10177 A AC 1.358 1.109 0.334 0.715 0.633 0.981 0.698 1.083 0.917 1.406 1.912
0.854 0.771 1.439 0.384 0.982 1.016 0.841 0.316 0.246 0.685 1.453 0.999 0.634 1.632 0.574 1.3
1.701 0.106 0.989 0.666 0.231 0.745 1.263 0.957 1.293 0.657 0.3 1.459 0.793 0.014 0.555 0.989
0.433 1.216 0.007 0.95 0.297 0.238 0.481 0.996 0.778 0.186 0.018 1.159 1.111 0.743 1.002 1.36
0.672 0.34 0.661 0.957 1.928 0.82 0.964 0.507 0.786 0.297 0.943 0.569 0.431 1.089 0.778 1.218
```

**Figure 8. Genotype file**

D) `laura_leroi.xls.` This file contains the information regarding the ID of the extreme phenotype individuals, needed to extract their epidemiological information from (`idsCVinformation.Rdata`).

E) `idsCVinformation.Rdata.` It contains the epidemiological information of the 2217 cases and controls from the SBC/EPICURO study, including the ones with extreme phenotypes. The header of the data is as follows:

*[**id**, region, age, gender, smoking status, CV folds, caco, casecontrol]*

F) `epicuro.samples.` Input file used in the imputation procedure containing the order and the IDs of the individual whose genotypes were imputed. It contains 1127 cases and 1090 controls, including the ones showing extreme phenotypes.

Following, Figure 9 shows graphically the file processing to extract the information needed for the rare variant analysis.

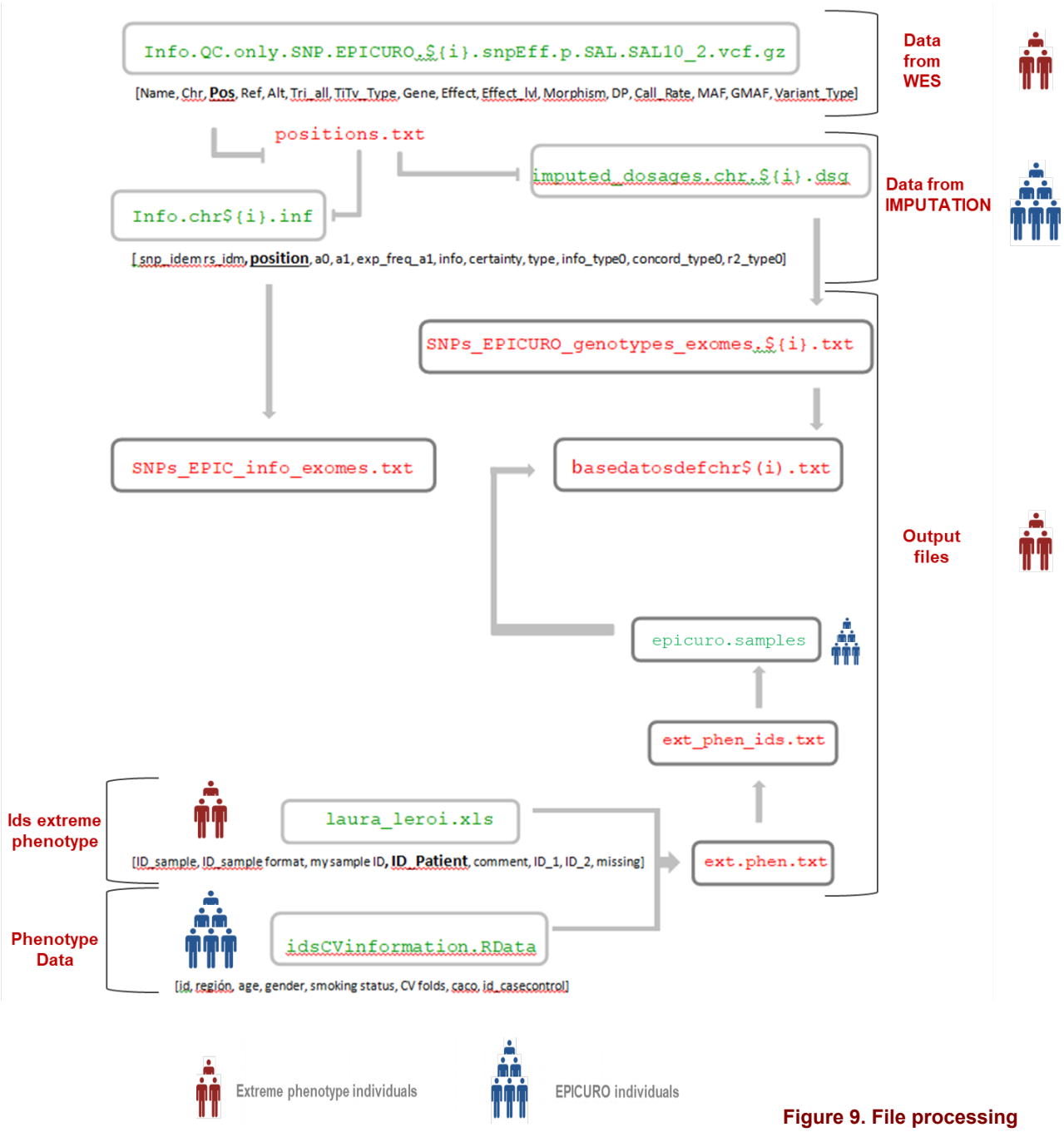


**Figure 9. File processing**

Since it was used an extreme phenotype design strategy and not all the individuals included in the SBC /EPICURO study had their exome sequenced, I needed to extract the genotypes of the sequenced SNPs after the imputation.

First, I extracted the positions of the SNPs from the file `Info.QC.only.SNP.EPICURO.$.snpEff.p.SAL.SAL10_2.vcf`, then I selected the rows containing that positions in the imputation files, and finally, I saved them into the file `position.txt`.
This was done in bash, using the following script:

```
#!/bin/bash
for i in {1..22};
do
      sed '1d' /local/comun/WES_extreme_EPICURO/EPICURO.$
{i}/info.QC.only.SNP.EPICURO.${i}.snpEff.p.SAL.SAL10_2.vcf.txt |cut -f3 >
/home/amfernandez/TFM_Alba/positions.txt
```

**Script 'to_extract_geno_info_exomes_FAST.sh'**

Then, I selected both the information and genotypes of the SNPs that were sequenced from the imputation files: `Info.$.chr.inf` and `imputed_dosages.$.chr.dsg.`

```
fgrep -w -f /home/amfernandez/TFM_Alba/positions.txt
/local/comun/EPICURO_imputed_by_MAF/chr${i}/Info.chr${i}.inf >
/home/amfernandez/TFM_Alba/SNPs_EPIC_info_exomes${i}.txt

fgrep -w -f /home/amfernandez/TFM_Alba/positions.txt
/local/comun/EPICURO_imputed_by_MAF/chr$i/imputed_dosages.chr${i}.dsg  >
/home/amfernandez/TFM_Alba/SNPs_EPICURO_genotypes_exomes${i}.txt
```

**Script 'to_extract_geno_info_exomes_FAST.sh'**

The output files were: 1) `SNPs_EPIC_info_exomes${i}.txt`, which contains the information of the WES SNPs, and 2) `SNPs_EPICURO_genotypes_exomes${i}.txt`, with the SNP genotypes.

Since the imputation files contain the genotypes for all the individuals of SBC/EPICURO, I needed to extract the genotypes of the individuals with extreme phenotypes. To do so, I used the file `epicuro.samples` to extract the corresponding columns in the genotype file, taking into account that the genotypes start in the field 6. Since the selected individuals were in the first 104 rows, the first 109 columns of `SNPs_EPICURO_genotypes_exomes${i}.txt` were saved as `basedatosdefchr${i}.txt,` using the code:

```
#!/bin/bash
for i in {1..22};
do
```

```
  echo $i
    cut -d " " -f 1-109
/home/amfernandez/tfm_Alba/SNPs_EPICURO_genotypes_exomes${i}.txt >
/home/amfernandez/tfm_Alba/basedatosdefchr${i}.txt
done
```

**Script 'basedatosdefchr.sh'**

Then, I extracted the epidemiological information for the subsample with WES data and saved it in the file `ext_phen_ids.txt`. To do so, I used `laura_leroi.xls` and `idsCVinformation.Rdata` as input files, and merged them by patient ID in R, using the following script:

```
install.packages('xlsx')
library('xlsx')
laura <-read.xlsx('/local/comun/TFM_Alba/___Laura Leroi.xlsx',1)
load('/local/comun/TFM_Alba/idsCVinformation.RData')
ext_phen <- merge(laura,all.ids.summary,by.x='ID_Patient',by.y='id')
write.table(ext_phen1,'/home/amfernandez/tfm_Alba/ext_phen1.txt',quote=FALSE,row
.names=FALSE,col.names=TRUE
write.table(ext_phen1$ID_Patient,'/home/amfernandez/tfm_Alba/ext_phen_ids.txt',q
uote=FALSE,row.names=FALSE,col.names=FALSE)
```

**Script 'ext_phen.R'**

## 3. 4. 1 Filtering of Variants

Only variants with a good imputation quality (info>0.3), polymorphic (i.e., whose genotype would vary in the set of cases and controls) and with minor allele frequency <1% were considered.

Rare variants filtering was done in R.

First, I loaded the necessary files

```
for(i in 1:22){
geno_exomes<-
read.table(file=paste0('/home/amfernandez/tfm_Alba/basedatosdefchr',i,'.txt'),st
ringsAsFactors=FALSE)
rownames(geno_exomes)=geno_exomes$V2
geno_exomes1<-geno_exomes[,c(6:109)]
info_exomes_impute <-
read.table(file=paste0('/home/amfernandez/tfm_Alba/SNPs_EPIC_info_exomes',i,
'.txt'), header=FALSE,stringsAsFactors=FALSE)
rownames(info_exomes_impute)=info_exomes_impute$V2
```

Then, I discarded the monomorphic:

```
geno_exomes1$monomorphic='NO'
```

```
for(j in 1: (dim(geno_exomes1)[1])){
  if (sum(geno_exomes1[j,1:104])/104 ==2) geno_exomes1$monomorphic[j]='YES2'
  if (sum(geno_exomes1[j,1:104])/104 ==0) geno_exomes1$monomorphic[j]='YES0'
  if ((min(geno_exomes1[j,1:104]) == max(geno_exomes1[j,1:104])) &
min(geno_exomes1[j,1:104]) ==1) geno_exomes1$monomorphic[j]='YES1'
}
no_monomorphic <-geno_exomes1[which(geno_exomes1$monomorphic == 'NO'),]
no_monomorphic$ID <- rownames(no_monomorphic)
no_monomorphic_data.frame <- data.frame(V2=rownames(no_monomorphic),
no_monomorphic)
```

After that, I filtered the variants by MAF:

```
info_exomes_impute$variant=rep('YES',length(info_exomes_impute$V6))
for(k in 1: (length(info_exomes_impute$V6))){
  if (info_exomes_impute$V6[k] <=0.05|info_exomes_impute$V6[k] >=0.95)
info_exomes_impute$variant[k] ='LOW'
  if (info_exomes_impute$V6[k] <=0.01|info_exomes_impute$V6[k] >=0.99)
info_exomes_impute$variant[k] ='RARE'
  if (info_exomes_impute$V6[k] >0.05 & info_exomes_impute$V6[k] <0.95)
info_exomes_impute$variant[k] ='COMMON'
}
table_MAF <- table(info_exomes_impute$variant
info_exomes_impute_rare<-
info_exomes_impute[which(info_exomes_impute$variant=='RARE'
info_exomes_impute_rare_no_monomorphic <-
info_exomes_impute_rare[rownames(info_exomes_impute_rare) %in%
no_monomorphic$ID,]
```

Finally, I filtered out those with bad imputation quality (info<0.3):

```
info_exomes_impute_rare_no_monomorphic_info <-
info_exomes_impute_rare_no_monomorphic[which(info_exomes_impute_rare_no_monomorp
hic$V7>=0.3),]
```

**Script 'scriptTFM'.R**

## 3. 5. Statistical Methods

The most popular approach in GWAS is to test each SNP individually and then prioritize those meeting a stringent significance level ($p$<0.05) after adjusting for multiple testing. However, individual SNP analysis in rare variant studies is seriously underpowered, due to the extremely low MAF or rare variants. Instead of testing each variant individually, novel statistical methods comprising aggregation tests have been developed recently. They evaluate cumulative effects of multiple genetic variants in a gene or region of interest, increasing the power when multiple variants in the group are associated with a given disease or trait [Seunggeung et al., 2014]

Aggregation tests involve two steps: first, to identify all rare variants within a sequenced (sub)-region (e.g., gene, regulatory region...) which passed the quality control filtering, and then, to test the joint effect of rare variants while adjusting for covariates [Thornton et al., 2015].
Numerous region or gene-based multi marker tests have been proposed in recent years [Zheng-Zheng et al., 2015]. The most commonly used are: the Burden tests [Morgenthaler and Thilly, 2007] [Li and Leal, 2008] [Madsen and Browning, 2009], Variance-Component Tests, including SKAT (Sequence Kernel Association Test) [Wu et al., 2010, 2011], and combined tests, which include SKAT-O (Sequence Kernel Association Test-Optimized) [Lee et al., 2012]

- Burden tests collapse rare variants in a genetic region into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants in the region. The score test for a weighted sum of genotypes has the form:

$$T = \sum_j w_j U_j,$$

where $w_j$ is a weight for SNP $j$ and $U_j$ is the score for SNP *j.* Weights can be used to upweight rare variants as for example in [Madsen and Browning, 2009].
Then, we test a null hypothesis with a single parameter H0: Uj =0, corresponding to fitting a simple logistic regression model.

- 

SKAT, on the other hand, is a weighted sum of individual score statistics. It aggregates the associations between variants and the phenotype through a kernel matrix and can allow for SNP-SNP interactions.

$$Q = \sum_j w_j^2 U_j^2$$

where $w_j$ is a weight and $U_j$ is the score statistic for the association between phenotype and variant *j.* The weight can be flexibly chosen.

- SKAT-O was proposed to test weighted averages of SKAT and burden tests. It has higher power in a wide range of settings, and is more robust than SKAT and the burden tests. The formula is:

$$Q_o(\rho) = (1-\rho)\left(\sum_{j=1}^{p} w_j^{skat} U_j^2\right) + \rho\left(\sum_{j=1}^{p} w_j^{burden} U_j\right)^2$$

When ρ = 0, this gives the SKAT test $Q_o = \sum_{j=1}^{p} w_j U_j^2$

When $\rho = 1$, it gives the burden test $Q_o = \left(\sum_{j=1}^{p} w_j U_j\right)^2$

[Seunggeun et al., 2012] [Voorman et al., 2011]

As it happens with other statistical methods, there is no single method that outperforms the rest in the analysis of rare variants. Their performance depends on the genetic architecture (e.g., effect directions and sizes) of each trait, which is unknown a priori. Burden tests are more powerful when most of the variants in a region are causal and the effects are in the same direction, whereas SKAT is more powerful when a large fraction of the variants in a region are non-causal or the effects of causal variants are in different directions [Seunggeun et al., 2012].

The following table summarizes the description, advantages and disadvantages of the different statistical methods:

| | *Description* | *Advantage* | *Disadvantage* |
|---|---|---|---|
| **BURDEN** | It collapses rare variants into genetic scores | They are powerful when a large proportion of variants are causal and effects are in the same direction | It loses power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants |
| **SKAT** | It tests variance of genetic Effects | It is powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants | It is less powerful than burden tests when most variants are causal and effects are in the same direction |
| **SKAT-O** | It combines burden and variance-component tests | It is more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait- decreasing variants | It can be slightly less powerful than burden or variance component tests if their assumptions are largely held is computationally intensive |

**Table 1. Comparative statistical methods**

Best powered test completely depends on the kind of causal variant.

## 3.6. Implementation: SeqMeta

For the implementation of the selected statistical tests we chose the Software Package **"SeqMeta"**: an R package for meta-analyzing region-based tests such as SKAT, SKAT-O, and burden test in the study of rare variants (https://rdrr.io/cran/seqMeta/).

The package can accommodate binary outcomes for unrelated individuals, as it is our case, and it provides functions for conditional analyses.
To implement each test, we needed to follow 2 steps: (1) to calculate summary statistics for each sequencing study and (2) to combine the summary statistics to perform gene-level association tests. [Tang et al., 2015]

### 3. 6. 1 prepScores

For the first step, we needed to run the function `prepScores` that returns an object that contains information referring to the genes, our unit of aggregation, with their respective MAFs in an .RData file as an output. This step is common for all methods, which are based on the output of this study level analysis to carry out its analysis.
The prepScores function needs the following arguments:

```
prepScores(Z=Z, formula=NullModel, SNPInfo=SNPInfo, data=pheno)
```

where:

`Z`: Is a matrix in which the columns correspond to the SNPs, while the rows correspond to the genotypes of the subjects.

`formula`: Is an object, adjusting for the possible covariates.

`SNPInfo`: Is a data frame with two columns: Name of the SNP and the gene in which it is located.

`data`: Contains the phenotypic information. It must have the same number of rows as Z and it must be in the same order (since they are the subjects).

Continuing with the previous script **Script 'scriptTFM'**, I adapted the files to create the different objects that the function `prepScores` needs.

First, I installed the package and loaded the SeqMeta library in R:

```
install.packages ('seqMeta')
library(seqMeta)
```

Then, I created the input objects needed to run the `prepScores` function.

- ✓ Get the genotype file, `Z`

We started with the variables defined above, in which we had applied the filtering of the variants and created the `Z` object. We kept the columns referring to the genotypes and transposed them to obtain the format that `Z` must have: individuals in rows and SNPs in columns.

```
Z <-no_monomorphic[no_monomorphic$ID%in%
rownames(info_exomes_impute_rare_no_monomorphic_info),]
```

```
Z<-Z[, c(1:104)]

Z1<-t(Z)
```

- The `formula`

Here the outcome is the case/control (`caco`). The model was adjusted whit the covariates region, age and gender

```
as.factor(caco)~as.factor(region)+ as.numeric(age)+ as.factor(gender)
```

- Get the SNP Info file, `SNPInfo`

To obtain the SNPs and their annotation in genes, we loaded the initial file `info.QC.only.SNP.EPICURO.',i,'.snpEff.p.SAL.SAL10_2.vcf.txt` that contains the name of the gene harboring each SNP and merged it with the variable in which we have saved only the SNPs that have passed the established filters. So we have a file that had two columns: the variant and the gene where the variant was located.

```
for(i in 1:22){

snpinfo <- read.table(file=
paste0('/local/comun/WES_extreme_EPICURO/EPICURO.',i,'/info.QC.only.SNP.EPICURO.
',i,'.snpEff.p.SAL.SAL10_2.vcf.txt'), header=TRUE,stringsAsFactors = FALSE) #

info_exomes_impute_Z <- info_exomes_impute[rownames(Z), ]

info_exomes_impute_Z←merge(info_exomes_impute_Z,snpinfo[,c('Pos','Gene')],by.x='
V3',by.y='pos')

rownames(info_exomes_impute_Z)=info_exomes_impute_Z$V2

SNPInfo<-info_exomes_impute_Z[,c('V2','Gene')]

colnames(SNPInfo)<-c("Name", "gene")}
```

- ✓ Get the `data` file

We uploaded our file with the phenotypic information

```
phenotypefile <- read.table(file= ('/home/amfernandez/tfm_Alba/ext_phen1.txt'),
sep = "", header = TRUE)

data<- phenotypefile
```

Before applying the function `prepScores`, I subset both Z and SNPinfo matrices into 2, one for genes harboring a single SNP and another one to harbor two or more SNPs. Here, I focused the analysis into the genes harboring 2 or more variants.

```
onevariant <- names(table(SNPInfo$gene)[table(SNPInfo$gene) < 2])
morethanonevariant<- names(table(SNPInfo$gene)[table(SNPInfo$gene)>= 2])
```

```
SNPInfo_unique <- SNPInfo[SNPInfo$gene %in% onevariant,]
SNPInfo_twoormore <- SNPInfo[SNPInfo$gene %in% morethanonevariant,]

Z1_unique <- Z1[,colnames(Z1) %in% SNPInfo_unique$Name]
Z1_twoormore <- Z1[,colnames(Z1) %in% SNPInfo_twoormore$Name]
```

Once all the necessary objects were obtain, we run the function `prepScore` to perform the study level analyses:

**Apply the `'prepScore'` function**

```
c1_twoormore<- prepScores(Z1_twoormore, as.factor(caco)~as.factor(region)
+as.numeric(age)+as.factor(gender), family = binomial(), SNPInfo =
SNPInfo_twoormore, data = data)
```

The results were stored in the `c1_twoormore` (for genes with two or more snps), which is needed as input for performing the rare variants association tests.

### 3. 6. 2 Burden test

We used the function `burdenMeta`, which takes as arguments:

- The result of the study level analyses `c1_twoormore`
- `wts`: we used a continuous weight function can be used to upweight rare variants `function(maf){1/(maf*(1-maf))}` [Madsen and Browning, 2009].
- the SNPInfo object `SNPInfo_twoormore`

```
out_burden.results=NULL

burden.results <- burdenMeta(c1_twoormore, wts = function(maf){1/(maf*(1-maf))},
SNPInfo = SNPInfo_twoormore)

out_burden.results <-rbind(out_burden.results, burden.results)
```

### 3. 6. 3 SKAT

We used the function `skatMeta`, which takes as arguments:

- The result of the study level analyses `c1_twoormore`.
- the SNPInfo object `SNPInfo_twoormore`

```
out_skat.results=NULL

skat.results <- skatMeta(c1_twoormore, SNPInfo = SNPInfo_twoormore)

out_skat.results <-rbind(out_skat.results, skat.results)
```

### 3. 6. 4 SKAT-O

Before running the SKAT-O analysis we needed to remove the SNP pairs in almost complete LD ($r^2$>0.95) in order to avoid numerical problems. Therefore, I programmed a function to: 1) calculate the linkage disequilibrium (LD) in those genes with two variants, 2) remove the SNP pairs that were in high LD $r^2$> 0.95 and 3) we created the `Z` object and the `SNPInfo` again, after discarding one SNP of each pair showing extremely high LD. To do so, I used the package genetics (https://cran.r-project.org). First, I converted the genotypes codified as dosages into the required format: D/D, D/I and I/I. Then, I created objects of the form genotype, and calculated the LD with the LD function.

```
genes_2SNPs=names(table(SNPInfo_twoormore$gene)
[which(table(SNPInfo_twoormore$gene)==2)])

calculateLD <- function(x){
  Z1_two=Z1_twoormore[,which(colnames(Z1_twoormore) %in%
SNPInfo_twoormore$Name[which(SNPInfo_twoormore$gene==x)])]
  Gdata<-ifelse(Z1_two=="2","D/D",ifelse(Z1_two=="1","D/I","I/I"))
  names(Gdata)
  snp1<-genotype(Gdata[,1])
  snp2<-genotype(Gdata[,2])
  LD_snps<-LD(snp1,snp2)
  r2<-LD_snps$r**2
}

genes_2SNPs_LD<-
data.frame(cbind(genes_2SNPs,unlist(lapply(genes_2SNPs,calculateLD))),stringsAsF
actors = FALSE)
names(genes_2SNPs_LD)=c('gene','r2')
genes_2SNPs_LD_toskip<-
genes_2SNPs_LD[which(as.numeric(genes_2SNPs_LD$r2)>0.95),]
SNPInfo_twoormore_skato <- SNPInfo_twoormore[-which(SNPInfo_twoormore$gene %in%
genes_2SNPs_LD_toskip$gene),]
```

**Apply the `'prepScore'` function**

```
c1_twoormore<- prepScores(Z1_twoormore[,which(colnames(Z1_twoormore) %in%
SNPInfo_twoormore_skato$Name)], as.factor(caco)~as.factor(region)
+as.numeric(age)+as.factor(gender), family = binomial(), SNPInfo =
SNPInfo_twoormore_skato, data = data)
```

Then, to carry out the analysis whit SKAT, I used the function of `skatOMeta`, which takes the following arguments:

- The object `c1_twoormore`, output of prepScores function.

- `Rho`: The values of ρ to be used in SKAT-O. By default is c(0,1), which computes SKAT and the burden test, and reports the minimum p-value adjusted for multiple testing.
- `skat.wts` and `burden.wts` which gives the weights to be used in SKAT and the burden test, respectively. In our case, the weigths used in the burden test comes from a Beta function of the MAF, with hyper parameters 1 and 25.
- The new SNPInfo object `SNPInfo_twoormore_skato`.

```
out_skato.results=NULL

skato.results <- skatOMeta(c1_twoormore, rho=seq(0,1,length=11), burden.wts =
function(maf){dbeta(maf,1,25)}, SNPInfo = SNPInfo_twoormore_skato, method =
"int")

out_skato.results <-rbind(out_skato.results, skato.results)
```

## 3. 7 Multiple testing correction

Once the p values were obtained after the application of the different statistical methods, it was necessary to apply a multiple tests correction. Benjamini-Hochberg (B-H) procedure [Ghosh, 2012] was chosen in order to avoid type I errors (false positives). Only those genes with an adjusted *p value <0.*05 were prioritized for further analyses.

```
out_skat.results$p_adj<- p.adjust(out_skat.results$p, method="BH")

skat.results_adjust <- out_skat.results[which(out_skat.results$p_adj<0.05),]

out_burden.results$p_adj<- p.adjust(out_burden.results$p, method="BH")

burden.results_adjust <- out_burden.results
[which(out_burden.results$nsnpsUsed!=0 & out_burden.results$p_adj<0.05),]

out_skato.results$p_adj<- p.adjust(out_skato.results$p, method="BH")

skato.results_adjust <- out_skato.results[which(out_skato.results$p_adj<0.05),]
```

## 3. 8 Manhattan plot

Once the adjusted p-value was calculated for each tested gene, I plotted them against gene's chromosomal position. These Manhattan plots were done with `'qqman'` (http://cran.r-project.org/web/packages/qqman/).

## 3. 9 Overlapping

I prioritized those genes that remained significant after multiple testing correction in the three association tests. To do so, I used the `VennDiagram` package (https://cran.r-project.org) and programmed a function to identify the significant genes identified with the three methods:

```
library('VennDiagram')
overlap <- function(y){
  overlap_totalbychr <- calculate.overlap(
  x = list(
  "skat" = skat.results_adjust$gene[which(skat.results_adjust$chr==y)],
  "burden" = burden.results_adjust$gene[which(burden.results_adjust$chr==y)],
  "skato" = skato.results_adjust$gene[which(skato.results_adjust$chr==y)]
)
)

  tabla<-data.frame(chr= y, skat=nrow(subset(skat.results_adjust,chr==y)),
  burden=nrow(subset(burden.results_adjust,chr==y)),
  skato=nrow(subset(skato.results_adjust,chr==y)),
  overlap=length(overlap_totalbychr$a5),
burden.skat=length(overlap_totalbychr$a2),
skat.skato=length(overlap_totalbychr$a4),
burden.skato=length(overlap_totalbychr$a6))

}
```

Then, I calculated the Spearman's correlation using the function in R `cor(data, method = "spearman")` between the ranking of genes obtained with the three tests: Burden, SKAT and SKAT-O.

## 3. 10 DisGeNET

I used the R package DisGeNET [Piñero et al., 2016] to find associations between the associated genes with neoplasms. DisGeNET (http://www.disgenet.org) is one of the largest available collections of genes and variants involved in human diseases and integrates data from expert curated repositories, GWAS catalogues, animal models and the scientific literature.

## 3. 11 Pathways Annotation

Genes that were identified by the three statistical methods were prioritized and then annotated in pathways using KEGG: a database resource (http://www.genome.jp/kegg/) that provides knowledge about genomes and their relationships to biological systems as well as their interactions with the environment [Aoki-Kinoshita et al., 2007].

First, we added 'hsa:' to the genes list, with significant genes (to find into *Homo sapiens*):

```
#!/bin/bash
```

```
sed 's/^/hsa:/g' gene_list_overlap > gene_list_overlap2
```

Then we added wget http://rest.kegg.jp/get/ to the genes list with hsa:

```
sed 's/^/wget http\:\/\/rest.kegg.jp\/get\//g' gene_list_overlap2 > nuevoscript.sh
```

We copy the results use the KEGG REST API to retrieve automatically KEGG ids genes, this will download the KEGG info, including pathway (example with one of the genes).

```
wget http://rest.kegg.jp/get/hsa:CLCNKB
```

Finally, to extract the pathways used the following code.

```
grep PATHWAY hsa* -A 20 | grep -P " hsa[0-9]" | sed 's/^hsa://g' > genes2pathways.txt
```

**Script 'genes2pathway_Alba.sh'**

Once I had the KEGG pathways, I selected the ones with the largest number of genes identified in the rare variant analysis (with a number of genes equal to or greater than 2) and colored them through the 'pathways' function of R.

```
library("pathview")
data(paths.hsa)
keggIDs <-
c("hsa01100","hsa05131","hsa04810","hsa04740","hsa04530","hsa04014","hsa05231","
hsa05205","hsa05200",
"hsa05152","hsa04922","hsa04666","hsa04514","hsa04510","hsa04144","hsa04024","hs
a04015","hsa04010","hsa03420","hsa05418","hsa03420")
pathGenes <-
c("ACADL","ARPC1A","CALML4","CHST6","CLDN6","CPSF3","CRLS1","CYP27B1","DIAPH1","
ERC1","ERCC1","HMGN1",
"LIG1","MPZ","MRPL1","NAPEPLD","NEDD4","OR2F1","OR2T8","PAK1","PLAUR","PMVK","RA
SGRP3","SIK1","SLC1A3","SLC22A5","SLC44A4","SLC9A3R1","SOCS5","TBX21","U2AF1L4")
read.table("~/top_pw_atleast2genes.txt", header=TRUE, row.names=1)
keggIDs <- rownames(pathgenes)
```

Then I plotted the KEGG pathways which where colored according to the number of SNPs per gene.

```
i=0
for (keggID in keggIDs[20]) {
   i=i+1; print(paste0(i,"/",length(keggIDs)," ", keggID))
pv.out <- tryCatch({pathview(gene.data=allGenes ,pathway.id=keggID,
species="hsa", out.suffix="overlap", kegg.native=TRUE, cpd.idtype="kegg",
gene.idtype="SYMBOL")}
}
```

**Script 'pathview.R'**

# 4. RESULTS AND DISCUSSION

## 4. 1 Summary

Table 2 shows a descriptive summary of the study population used in this study. The categorization of the tobacco habit was made by separating smokers and non-smokers, and within these, was stratified into two levels: those who smoked 20-40 cigarettes a day and those who smoked > 40.

The controls were all males, older than 70 and heavy smokers. On the contrary, ~37% of the cases were male, half of them were < 70 and all of them were non-smokers.

| | Control N=36 | Case N=68 |
|---|---|---|
| gender: | | |
| Male | 36 (100%) | 25 (36.8%) |
| Female | 0 (0.00%) | 43 (63.2%) |
| agecat: | | |
| ≤50 | 0 (0.00%) | 8 (11.8%) |
| 51-69 | 0 (0.00%) | 26 (38.2%) |
| ≥70 | 36 (100%) | 34 (50.0%) |
| smoking_status: | | |
| Non-Smoker | 0 (0.00%) | 68 (100%) |
| 20-40 | 26 (72.2%) | 0 (0.00%) |
| 40+ | 10 (27.8%) | 0 (0.00%) |

**Table 2. Summary descriptive by Case-Control**

## 4. 2 SNPs data

Figure 10 displays the total number of SNPs that we handled during each step in the process of variant calling, quality control, imputation and filtering of rare variant.

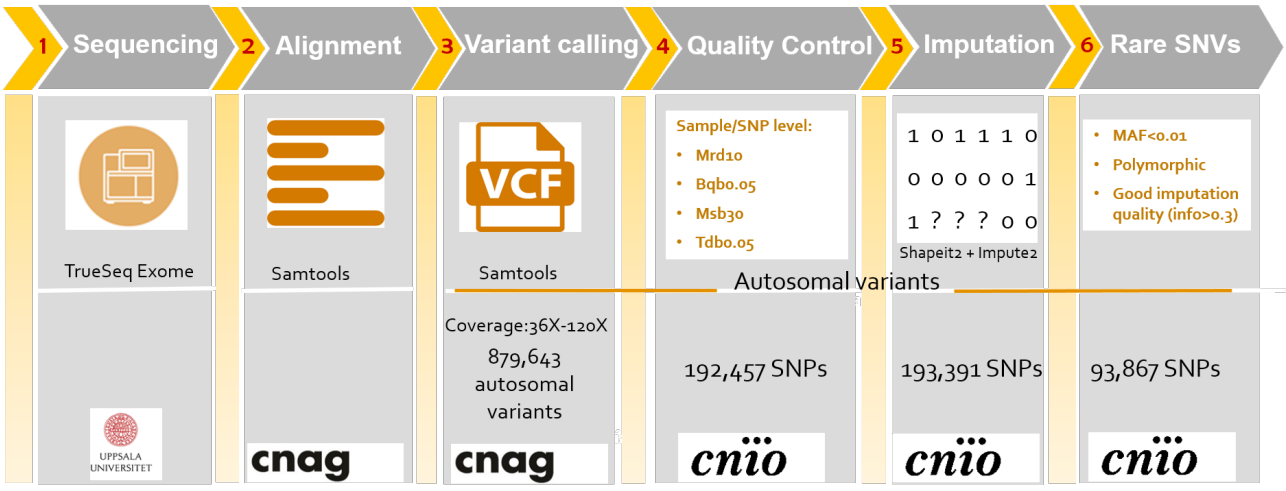


**Figure 10. Workflow with SNPs numbers**

After the variant calling, a total of 879,643 autosomal variants were available. Once, the INDELS were filtered out and the quality control procedure was applied to ensure the quality of the genotypes, a total of 192,457 SNPs remained. After imputing the missing genotypes, and due to the presence of multi-allelic SNPs, a total number of 193,391 SNPs were available. Then, 93,867 rare variants (MAF<0.01) with good imputation quality (info> 0.3), which represent the 48.54% of the initial number of SNPs, were selected for the gene-based association analyses.
Supplementary Table 1' (in the annex) shows the number of SNPs that remained after each filtering step by chromosome. Between 33.47% and 60.46% of the initial SNPs (after imputation) in chromosomes 18 and 19 were polymorphic rare variants with good imputation quality.

## 4. 3 Gene-base analysis

Three aggregation tests on a gene basis were considered in this study: Burden, SKAT and SKAT-O. In contrast to individual tests, aggregation tests evaluate cumulative effects of multiple genetic variants in a gene or region, increasing power when multiple variants in the group are associated with the disease of interest. Table 3 shows the summary statistic of the cumulative MAF of the variants used as input for the three aggregation tests.

Cumulative MAF ranges from a minimum of 0.0096 to a maximum of 0.32, with a median of 0.019.

| | *Minimum* | *1st Quartile* | *Median* | *Mean* | *3rd Quartile* | *Maximum* |
|---|---|---|---|---|---|---|
| **cMAF** | 0.009615 | 0.01442 | 0.01923 | 0.03644 | 0.03606 | 0.3221 |

**Table 3. cMAF summary statistic**

## 4. 4 Burden results

### 4. 4. 1 burdenMeta function output

The output of the test was saved in an object with the following information: *chr* (chromosome in which the gene is found), *gene* (the gene name), *p* (the p-value from Burden test), *beta* (a parameter to report estimated effects), *se* (the standard error of beta), *cmafTotal* (the total cumulative minor allele frequency), *cmafUsed* (cumulative minor allele frequency used), *nsnpsTotal, nsnpsUsed* and *nmiss* (the number of missing SNPs).
Table 4 shows the output from the burden test:

| chr | gene | p | beta | se | cmafTotal | cmafUsed | nsnpsTotal | nsnpsUsed | nmiss |
|---|---|---|---|---|---|---|---|---|---|
| 1 | LINC00115 | 0.3500 | -2.0718 | 2.2168 | 0.0288 | 0.0048 | 2 | 1 | 0 |
| 1 | LINC01128 | 0.9756 | 1.0073 | 32.9942 | 0.0529 | 0.0096 | 5 | 2 | 0 |
| 1 | SAMD11 | 0.7521 | -0.4049 | 1.2818 | 0.1298 | 0.0433 | 10 | 8 | 0 |
| 1 | NOC2L | 0.8743 | -0.1877 | 1.1865 | 0.5577 | 0.0529 | 18 | 9 | 0 |
| 1 | KLHL17 | 0.1546 | -1.3518 | 0.9496 | 0.5288 | 0.0577 | 20 | 10 | 0 |
| 1 | C1orf170 | 0.5012 | 0.8891 | 1.3218 | 0.3365 | 0.0337 | 12 | 6 | 0 |

**Table 4. burdenMeta function results**

Once the multiple test correction was applied using Benjamini-Hochberg method, and after considering as significant only those genes with an adjusted p-value <0.05, a total of **184 genes** were remained.

Figure 11 represents the Manhattan plot obtained with the burden test. The 184 significant genes (adjusted *p*<0.05) are highlighted in green. In addition, the top genes according to their p-value were annotated.

### 4. 4. 2 Manhattan plot of burden test

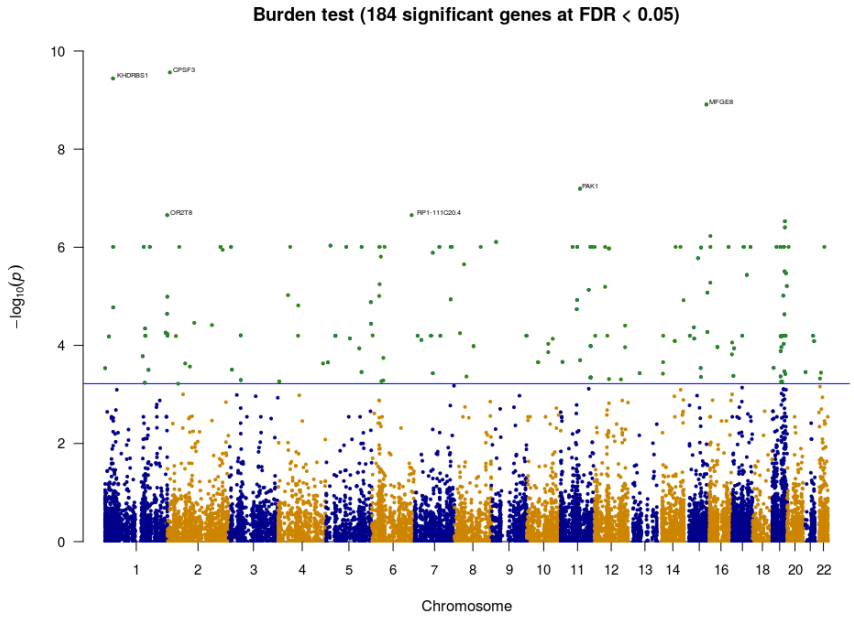


**Figure 11. Manhattan plot of Burden results**

## 4. 5 SKAT results

### 4. 5. 1 skatMeta function output

Table 5 shows the output return with the SKAT method:

| chr | gene | p | Qmeta | cmaf | nmiss | nsnps |
|---|---|---|---|---|---|---|
| 1 | LINC00115 | 0.6086 | 94.4163 | 0.0288 | 0 | 2 |
| 1 | LINC01128 | 0.2073 | 255.9366 | 0.0529 | 0 | 5 |
| 1 | SAMD11 | 0.2911 | 565.6360 | 0.1298 | 0 | 10 |
| 1 | NOC2L | 0.3674 | 589.0266 | 0.5577 | 0 | 18 |
| 1 | KLHL17 | 0.1589 | 911.1544 | 0.5288 | 0 | 20 |
| 1 | C1orf170 | 0.7896 | 270.8077 | 0.3365 | 0 | 12 |

**Table 5. skatMeta function results**

This output format is similar to that obtained with burdenMeta. As new parameters include *Qmeta*, the statistic score of SKAT.

The number of significant genes (with an adjusted *p*-value< 0.05) obtained by skatMeta was **169**. The Manhattan plot for this output is shown in figure 12.

### 4. 5. 2 Manhattan plot of SKAT test

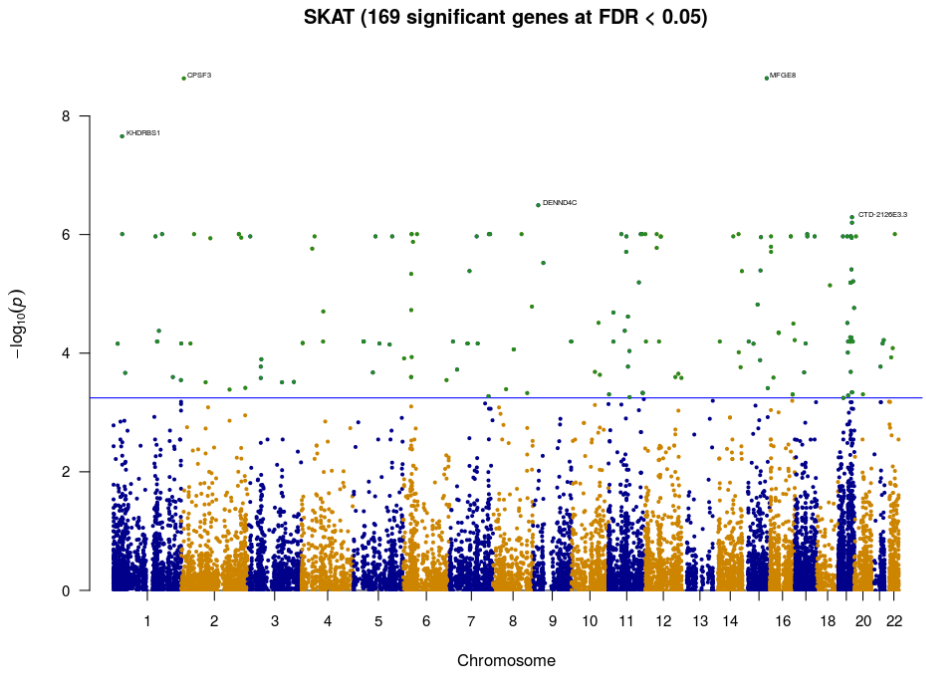


**Figure 12. Manhattan plot of SKAT results**

## 4. 6 SKAT-O results

### 4. 6. 1 skatOMeta function output

Finally, the results obtained from the skatOmeta function are shown in table 6:

| chr | gene | p | pmin | rho | cmaf | nmiss | nsnps | errflag |
|---|---|---|---|---|---|---|---|---|
| 1 | LINC00115 | 0.7314 | 0.6079 | 0.0 | 0.0288 | 0 | 2 | 0 |
| 1 | LINC01128 | 0.1898 | 0.1625 | 1.0 | 0.0529 | 0 | 5 | 0 |
| 1 | SAMD11 | 0.4315 | 0.2958 | 0.0 | 0.1298 | 0 | 10 | 0 |
| 1 | NOC2L | 0.5202 | 0.3707 | 0.0 | 0.5577 | 0 | 18 | 0 |
| 1 | KLHL17 | 0.0884 | 0.0762 | 1.0 | 0.5288 | 0 | 20 | 0 |
| 1 | C1orf170 | 0.9345 | 0.7905 | 0.0 | 0.3365 | 0 | 12 | 0 |

**Table 6. skatOMeta function results**

The output of SKAT-O is very similar to that of skatMeta, but with two additional parameters *pmin*, (the minimum p-value among the tests), and *rho,* a parameter between 0 and 1, which is also

calculated in the function and represents the weight given to the Burden test (being 1-rho the weight given to SKAT).

Similarly to to Burden or SKAT approaches, only those genes with adjusted p-values below 0.05 were considered as significant. SKAT-O was the method with the highest number of significant genes returned, with a total of **197 genes**.

The Manhattan plot with the genes identified by this method is shown in Figure 13.

### 4. 6. 2 Manhattan plot of SKAT- O test

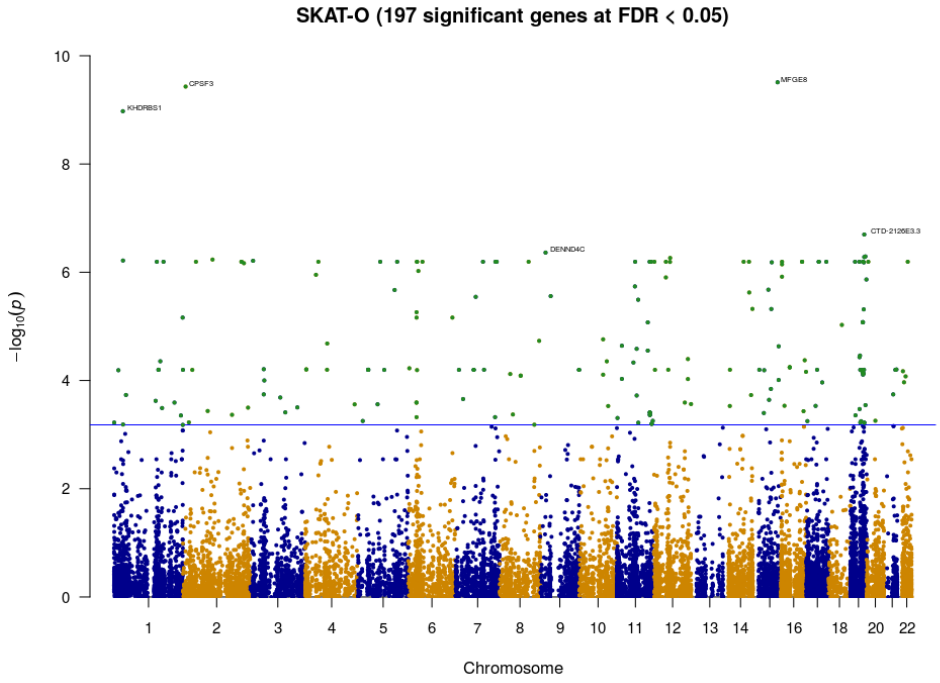


**Figure 13. Manhattan plot of SKAT-O results**

## 4. 7 Gene prioritization

The performance of aggregation methods, which are based on varying assumptions about the underlying genetic model, depends on the true disease model. Because the true disease model is unknown, we decided to prioritize those genes that were significant with the three methods. Figure 14 shows the Venn diagram representing the overlapping between the genes obtained with each aggregation method. A good overlapping among the results of the three tests was obtained, with 119 genes identified by the 3 methods. SKAT and SKAT-O showed a very good overlapping (only 4 genes were detected by SKAT and not by SKAT-O, and 8 were detected with SKAT-O but not with SKAT). On the other hand, Burden and SKAT showed the worst overlapping, since Burden test identified 41 genes that were not identified by SKAT-O. Burden test was the method that identified the largest number of genes that were not identified by the remaining two methods (39 genes).

The rankings of genes obtained with the three methods also showed a good agreement: Spearman's correlation of the ranking obtained with the Burden test and that obtained with SKAT was the lowest one (0.744). However, the ones obtained with Burden and SKAT-O, and those with SKAT and SKAT-O were 0.82 and 0.89, respectively. These figures may suggest that, in general, the underlying genetic model is closer to the one assumed by SKAT (variants with different directions and effects) than the one assumed by the Burden test.

Supplementary Table 2' shows the number of significan genes obtained by chromosome, for each method, and for the overlap.

Supplementary Table 3' shows the *p values* of the 3 methods for each of the 119 significant overlapping genes, as well as a short description.

Below, Table 7 show the top 10 of the statistically significant genes.

| *Symbol* | *pval_Burden* | *pval_SKAT* | *pval_SKATO* | *Description* |
|---|---|---|---|---|
| **CPSF3** | 2.72E-010 | 2.33E-009 | 3.71E-010 | cleavage and polyadenylation specific factor 3 |
| **MFGE8** | 1.23E-009 | 2.31E-009 | 3.08E-010 | milk fat globule-EGF factor 8 protein |
| **KHDRBS1** | 3.61E-010 | 2.20E-008 | 1.06E-009 | KH RNA binding domain containing, signal transduction associated 1 |
| **CTD-2126E3.3** | 2.95E-007 | 5.09E-007 | 2.00E-007 | <NA> |
| **DENND4C** | 7.78E-007 | 3.20E-007 | 4.33E-007 | DENN domain containing 4C |
| **ZNF473** | 3.92E-007 | 6.30E-007 | 5.23E-007 | zinc finger protein 473 |
| **ZBTB8B** | 9.82E-007 | 9.82E-007 | 6.07E-007 | zinc finger and BTB domain containing 8B |
| **TOR3A** | 9.82E-007 | 9.82E-007 | 6.39E-007 | torsin family 3 member A |
| **RP11-328C8.2** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |

**Table 7. *P-value* and description of top10 significant genes**

### 4. 7. 1 VennDiagram

The graphic representation of this overlap is shown in Figure 14.

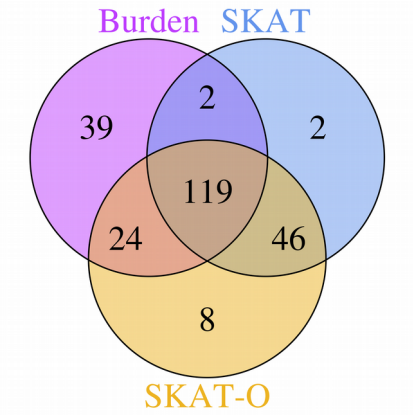


**Figure 14. Venn Diagram of significant genes by the three methods**

Supplementary Table 3' provides detailed results of the significant genes by chromosome, obtained by each method, as well as those obtained by the total overlapping, or compared by pairs of methods.

### 4. 7. 2 Gene disease-associations.

Twelve genes out of the 119 prioritized in the rare variants association analysis were previously associated with other neoplasms according disgenetGene to function (see Table 8). Interestingly, two of them *LIG1* and *ERCC1*were previously associated with bladder neoplasms. Furthermore, genes like *ANXA3* or *PLAUR*, were previously associated with prostate cancer, and others like *PAK1* or *IL32* were linked to renal cell and kidney carcinomas, respectively.

| Gene | Disease |
|---|---|
| ***PMVK*** | *Malignant neoplasm of skin, Skin Neoplasms* |
| ***HOXD9*** | *Colorectal Neoplasms, Mucinous Adenocarcinoma, ovarian neoplasm* |
| ***ANXA3*** | *Prostatic Neoplasms, ovarian neoplasm* |
| ***HOXA2*** | *Stomach Neoplasms* |
| ***PAK1*** | *Mammary Neoplasms, Renal Cell Carcinoma* |
| ***MFGE8*** | *Mammary Neoplasms, Liver Neoplasms, Experimental* |
| ***IL32*** | *Colonic Neoplasms, Sezary Syndrome, Kidney Neoplasm, Stomach Neoplasms* |
| ***BCL7C*** | *Ependymoma* |
| ***PLAUR*** | *Neoplasm Metastasis, Prostatic Neoplasms, Neoplasm Invasiveness* |
| ***ERCC1*** | *Non-Small Cell Lung Carcinoma, Stomach Neoplasms, Neoplasm Metastasis, melanoma, Testicular Neoplasms, Uterine Cervical Neoplasm, Neoplasms, Germ Cell and Embryonal* |

**Table 8. Identified genes associated with other neoplasm according to Disgenet**

### 4. 7. 3 Pathways annotation

The most frequent, annotated pathways of the 119 significant genes are presented in Table 9.

| ID_pathway | Pathway | N. of genes |
|---|---|---|
| hsa01100 | *Metabolic pathways* | 4 |
| hsa05131 | *Shigellosis* | 3 |
| hsa04810 | *Regulation of actin cytoskeleton* | 3 |
| hsa04740 | *Olfactory transduction* | 3 |
| hsa04530 | *Tight junction* | 3 |
| hsa04014 | *Ras signaling pathway* | 3 |
| hsa05231 | *Choline metabolism in cancer* | 2 |
| hsa05205 | *Proteoglycans in cancer* | 2 |
| hsa05200 | *Pathways in cancer* | 2 |
| hsa05152 | *Tuberculosis* | 2 |
| hsa04922 | *Glucagon signaling pathway* | 2 |
| hsa04666 | *Fc gamma R-mediated phagocytosis* | 2 |
| hsa04514 | *Cell adhesion molecules (CAMs)* | 2 |
| hsa04510 | *Focal adhesion* | 2 |
| hsa04144 | *Endocytosis* | 2 |
| hsa04024 | *cAMP signaling pathway* | 2 |
| hsa04015 | *Rap1 signaling pathway* | 2 |
| hsa04010 | *MAPK signaling pathway* | 2 |
| hsa03420 | *Nucleotide excision repair* | 2 |
| hsa05418 | *Fluid shear stress and atherosclerosis* | 2 |

**Table 9. Most common pathways of significant genes**

According to KEGG pathways, the most interesting pathways are:

- "Nucleotide excision repair", which has other BC susceptibility genes as *XPC* and *XPD*, previously found in a pooled association analysis [Stern et al., 2009]
- "Regulation of actin cytoskeleton", which has other BC susceptibility genes as *FGFR3*, previously found in GWAS. [López de Maturana et al., 2017]
- "Tight junction" which has 3 posible susceptibility genes found in our analysis. In particular, *Claudin 6* gene is found expressed in several tumour cells, and the methylation of this gene may be involved in oesophageal tumorigenesis.

Figure 15 shows the *nucleotide pathway excision repair*, where we can appreciate the steps where the *ERCC* intervenes on this route.

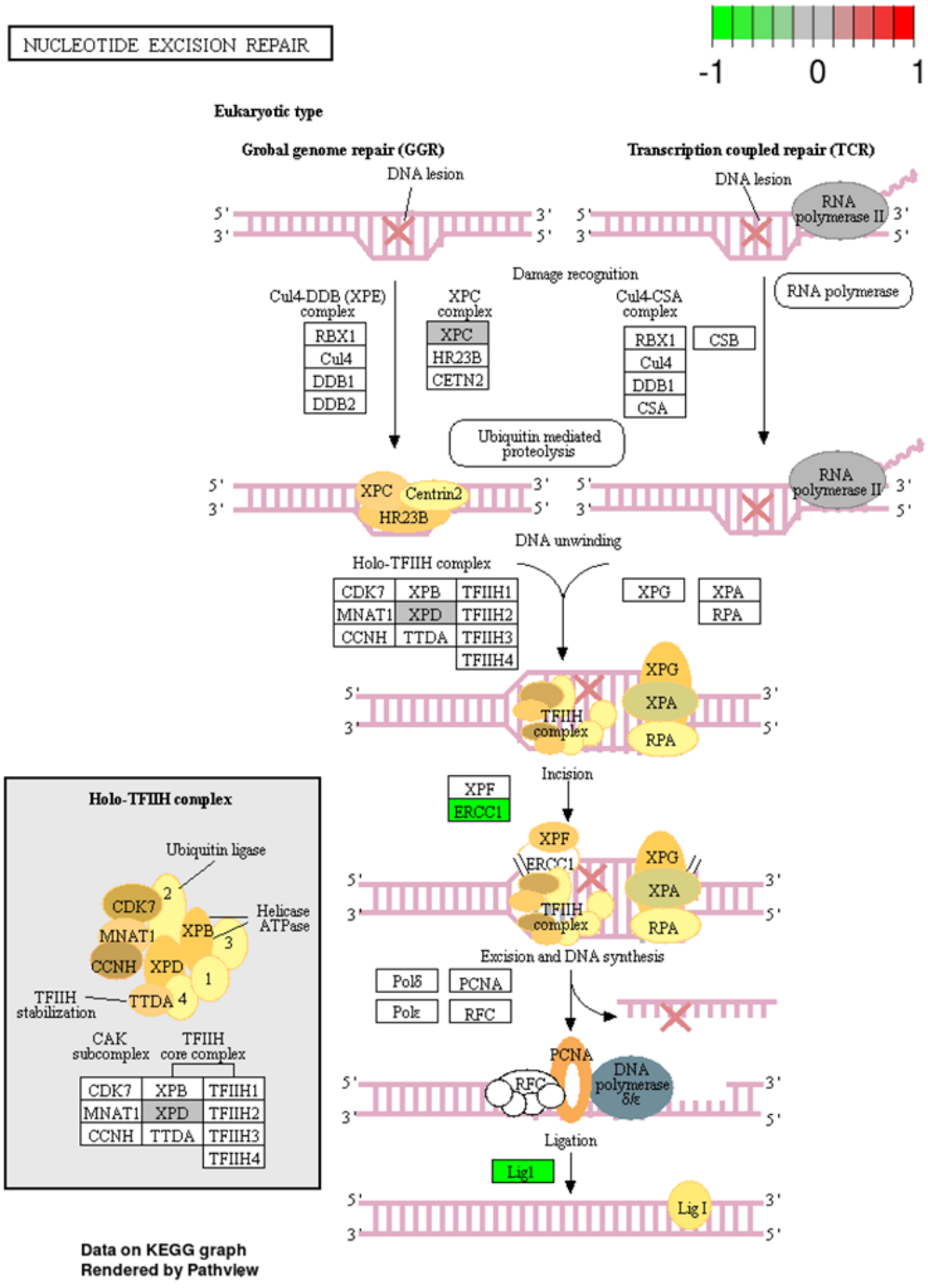


**Figure 15. Nucleotide excision repair pathway**

### 4. 7. 4 Variants in the 119 significant genes

The number of variants included in the 119 significant genes obtained from the superposition of the 3 methods was 510 SNPs.

Figure 16 shows, the number of SNPs that overlap in each of the 22 chromosomes.

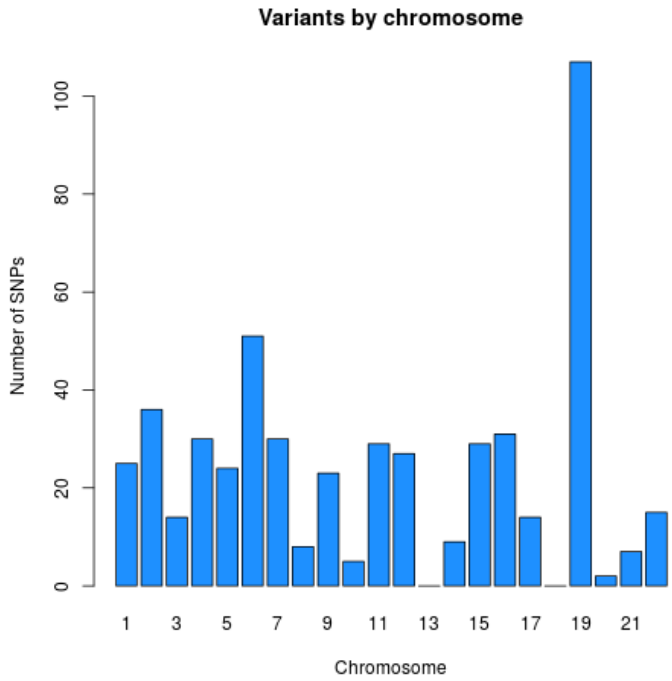


**Figure 16. Number of SNPs by Chromosome**

### 4. 7. 5 Variant effect and impact

Table 10 shows the description of both the SNP effects and their impact, considering the initial number of SNPs from which we started (93,867), as well as those found in the significant genes (510) found with the three methods. Table 10 also includes a description of the effects. Interestingly, none of the variants have a high effect among the ones that are harbored in the significant genes. However, there are twenty-four with a moderate effect.

| Effect | Description | Impact | Variants | Variants in sig. genes |
|---|---|---|---|---|
| *Downstream* | Downstream of a gene (5 Kb) | Modifier | 43616 | 254 |
| *Exon* | The variant hits an exon (from a non-coding transcript) or a retained intron | Modifier | 14644 | 93 |
| *Intergenic* | The variant is in an intergenic region | Modifier | 483 | 0 |
| *Intragenic* | The variant hits a gene but no transcripts within the gene | Modifier | 5 | 0 |
| *Intron* | The variant hits an intron (hits no exon in the transcript) | Modifier | 22573 | 126 |
| *Upstream* | Upstream of a gene (5 Kb) | Modifier | 251 | 1 |
| *3'-UTR* | Variant hits 3'-UTR region | Modifier | 223 | 0 |
| *5'-UTR* | Variant hits 5'-UTR region | Modifier | 86 | 1 |
| *Non-synonymous start* | Variant causes start codon to be mutated into another start codon (the new codon produces a different amino acid) | Low | 1 | 0 |
| *Splice site region* | A sequence variant in which a change has occurred within the region of the splice site (either within 1-3 bases of the exon or 3-8 bases of the intron) | Low | 106 | 1 |
| *Start gained* | A variant in 5'-UTR region produces a 3 bp sequence that can be a start codon | Low | 88 | 0 |
| *Synonymous coding* | Variant causes a codon that produces the same amino acid | Low | 4279 | 10 |
| *Synonymous stop* | Variant causes stop codon to be mutated into another stop codon | Low | 2 | 0 |
| *Non-synonymous coding* | Variant causes a codon that produces a different amino acid | Moderate | 7303 | 24 |
| *Start lost* | Variant causes start codon to be mutated into a non-start codon | High | 11 | 0 |
| *Stop gained* | Variant causes a stop codon | High | 194 | 0 |
| *Stop lost* | Variant causes stop codon to be mutated into a non-stop codon | High | 2 | 0 |
| | | **Total** | **93867** | **510** |

**Table 10. Effect and impact of variants**

# 5. SUMMARY

To our knowledge, this work represents the first approach to decipher the role that rare variants may have in the genetic susceptibility to BC.

I applied the three most commonly used aggregation tests: Burden test, SKAT and SKAT-O to identify genes harboring rare variants which are potentially associated with BC, by using a WES-based approach, where the individuals were selected using an extreme phenotype design.

Although some of the associations detected may be false positive results, especially due to the limited sample size, the large number of significant genes obtained with the three methods after multiple testing correction (119 genes) suggests that rare variants are likely to play a role in BC development. Only twelve genes were previously associated with any neoplasms, and therefore, most of the significant genes are novel susceptibility genes for BC. However, two genes identified in this approach (*LIG1* and *ERCC1* in the "Nucleotide excision repair" pathway) were reported previously as associated with BC, which supports these results. Another interesting pathway ("Regulation of actin cytoskeleton") has three novel BC susceptibility genes identified in this study in addition to the already known BC susceptibility gene *FGFR3*. These results add evidence to the participation of these two pathways through both common and rare variants in the BC development. Another novel and interesting pathway identified here is "Tight junction", with three significant genes. Among them, *Claudin 6* gene is the most relevant one, as its methylation has been associated with tumorigenesis. However, as it happens with the rest of the novel BC susceptibility genes identified in this study, their association with BC has to be validated in an independent and larger population, through a targeted sequencing strategy.

# 6. CONCLUSIONS

The conclusions of this work are:

- There is a good agreement between the ranking of genes according the their p-value which were obtained with the three methods.
- The large number of genes selected by the three methods (119) suggests that rare inherited coding variants across many genes contribute to bladder cancer genetic susceptibility.
- Rare variants associated with bladder cancer susceptibility are both in pathways already identified through GWAS and in novel pathways.

# 7. REFERENCES


Aoki-Kinoshita KF, Kanehisa M. 2007. Gene annotation and pathway mapping in KEGG. *Methods in molecular biology; 396:71-91.*

Benjamini Y, Hochberg J. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B. 1995; V.57 (No. 1):289–300.*

Burger M, James W, Catto F, Guido D, Grossman B, Herr H, Karakiewicz P, Kassouf W, Kiemeney LA , La Vecchia C, and others. 2013. Epidemiology and Risk Factors of Urothelial Bladder Cancer. *European Urology; (63): 234-241*

Chen H, Lumley T, Brody J , Heard-Costa NL, Fox CS , Cupples LA, Dupu J. 2014. Sequence Kernel Association Test for Survival Traits. *Genetic Epidemiology. 38(3): 191–197.*

Chen H. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. (https://cran.r-project.org/package=VennDiagram)

Cohen SM, Shiari T, Steineck G. 2000. Epidemiology and etiology of premalignant and malignant urothelial changes. *Scandinavian journal of urology and nephrology. Supplementum. (205):105-15.*

Epstein JI, Amin MB, Reuter VR, Mostofi FK. 1998. The World Health Organization/International Society of Urological Pathology consensus classification of urothelial (transitional cell) neoplasm of the urinary bladder. Bladder Consensus Conference Committee. *The American Journal of Surgical Pathology. 22(12):1435-48.*

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritabilty and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics. 11(6):446-50.*

Ferlay J, Steliarova-Foucher E, Lortet-Tieulent J, Rosso S, Coebergh JW, Comber H, Forman D, Bray F. 2013. Cancer incidence and mortality patterns in Europe: estimates for 40 countries in 2012. *European Journal of Cancer 49, 1374-1403.*

Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M , Parkin DM, Forman D, Bray F, Guillaume L and others. 2014 . Epidemiology of and risk factors for bladder cancer and for urothelial tumors. *International Journal of Cancer Rev Prat 64, 1372-4, 1378-80.*

Figueroa JD, Middlebrooks CD, Banday AR, Ye Y, Garcia-Closas M, Chatterjee N, Koutros S, Kiemeney LA, Rafnar T, Bishop T, and others. 2016. Identification of a novel susceptibility locus at 13q34 and refinement of the 20p12.2 region as a multi-signal locus associated with bladder cancer risk in individuals of European ancestry. *Human Molecular Genetics; 25(6):1203-14.*

García-Closas M, Malats N, Silverman D, Dosemeci M, Kogevinas M, Hein DW, Tardón A, Serra C, Carrato A, García-Closas R, and others. 2005. NAT2 slow acetylation and GSTM1 null genotypes increase bladder cancer risk: results from the Spanish Bladder Cancer Study and meta-analyses. *Lancet, 366(9486):649-59.*

Ghosh D. 2012. Incorporating the empirical null hypothesis into the Benjamini-Hochberg procedure .*Statistical Application in Genetic and Molecular Biology. 26; 11(4).*

Gibson G. 2015. Rare and common variants: twenty arguments. *Nature reviews. Genetics 13(2):135-45.*

GLOBOCAN 2012: estimated cancer incidence, mortality, and prevalence worldwide in 2012. International Agency for Research on Cancer. *Web site.* http://globocan.iarc.fr

Guillaume L, Guy L. 2014. [Epidemiology of and risk factors for bladder cancer and for urothelial tumors]. *La Revue du Practician: 64(10):1372-4, 1378-80.*

Jankovic S, Radosavljevic V. 2007. Risk factors for bladder cáncer. *Tumori 93(1):4-12.*

Jiekun X, Lei G, Leming S. 2012. Next-generation sequencing in the clinic: Promises and challenges. *Cancer letters 340(2).*

Leal J, Luengo-Fernández R, Sullivan R, Witjes AJ. 2016. Economic Burden of Bladder Cancer Across the European Union. *European Urology 69, 438-447.*

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *American Journal Human Genetics; 83(3):311-21.*

López de Maturana E, Chanok SJ, Picornell AC, Rothman N, Herránz J, Calle ML, García-Closas M, Marenne G, Brand A, Tardón A, and others. 2014. Whole Genome Prediction of Bladder Cancer Risk With the Bayesian LASSO. *Genetic Epidemiology 38(5):467-76.*

López de Maturana E, Malats N. 2017. Genetic Testing, Genetic Variation, and Genetic Susceptibility. *revised*

Low SK, Takahashi A, Mushiroda T, Kubo M. 2018. Genome-Wide Association Study: A Useful Tool to Identify Common Genetic Variants Associated with Drug Toxicity and Efficacy in Cancer Pharmacogenomics. *American Association for Cancer Research. 20(10); 2541-52.*

Luo W. 2013. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*, 29(14), (10.1093/bioinformatics/btt285)

Madsen BE, Browning SR. 2009. A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *PLoS Genet 5(2).*

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos J, Cardon LR, Chakravarti A, and others. 2009. Finding the missing heritability of complex diseases, *Nature 461(7265): 747–753.*

Morgenthaler S, Thilly WG. 2007. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanism of Mutagenesis. 615; 28-56*

Moyer VA, U.S. Preventive Services Task Force. 2011. Screening for bladder cancer: U.S. Preventive Services Task Force recommendation statement. *Annals of internal medicine; 155(4):246-51*

Ognenovski M, Renauer R, Gensterblum E, Kotter I, Xenitidis T, Henes JC, Casali B, Salvarani C, Direskeneli H, Kaufman KM, and others. 2016. Whole Exome Sequencing Identifies Rare-Coding Variants in Behcet´s Disease. *Arhritis & rheumatology ; 68(5) :1272-80.*

Ouzzanne A, Roupret M, Leon P, Yates DR, Colin P. 2014. Epidemiology and risk factors of upper urinary tract tumors: literature review for the yearly scientific report of the French National Association of Urology. *Progrès en urologie : journal de l'association française d'urologie et de la Société française d'urologie, 24(15) :966-76.*

Pan W, Kim J, Zhang Y, Shen X, Wei P. 2014. A Powerful and Adaptive Association Test for Rare Variants. *Genetics, Vol. 197, 1081–1095*

Pelucchi C, Bosetti C, Negri E, Malvezzi M, La Vecchia C. 2006. Mechanisms of disease: the epidemiology of bladder cancer*; Nature clinical practice. Urology. 3(6):327-40.*

Piñero J, Bravo A, Queralt-Rosinach N, Gutierrez-Sacristan A, Deu-Pons J, Centeno E, García-García J, Sanz F, Furlong LI. 2016. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research, 2017, Vol. 45, Database issue*

Rabbani B, Tekin M, Mahdieh N. 2014. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics; 59(1):5-15.*

Radosavljevic V, Belojevic G. 2014. Shortcomings in bladder cancer etiology research and a model for its prevention. *Tumori, 100: 1-8.*

Rothman N, Garcia-Closas M, Chatterjee N, Malats N, Wu X, Figueroa JD, Real FX, Van Den Berg D, Matullo G, Baris D, and others. 2010. A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. Nature Genetics *42(11):978-84.*

Samanic C, Kogevinas M, Dosemeci M, Malats N, Real FX, Garcia-Closas M, Serra C, Carrato A, García-Closas R, Sala M, and others. 2016. Smoking and bladder cancer in Spain: effects of tobacco type, timing, environmental tobacco smoke, and gender. *Cancer Epidemiology Biomarkers and Prevention 5(7):1348-54.*

Seunggeung L, Goncalo RA, Boehnke M, Xihong L. 2014. Rare-Variant Association Analysis: Study Designs and Statistical Test. *The American Journal of Human Genetics 95, 5-23.*

Seunggeun L, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. 2012. Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies. *The American Journal of Human Genetics 91, 224–237.*

Siegel RL, Miller KD, Jemal A. 2017. Cancer Statistics, 2017. *Cancer Journal for Clinicians; 67:7–30.*

Stern MC, Lin J, Figueroa JD, Kelsey KT, Kiltie AE, Yuan JM, Matullo G, Fletcher T, Benhamou S, Taylor JA. 2009. Polymorphisms in DNA repair genes, smoking, and bladder cancer risk: findings from the international consortium of bladder cáncer. *Cancer Research. 1; 69(17)*

Zeng-Zeng T, Lin DY. 2015. Meta-analysis for Discovering Rare-Variant Associations: Statistical Methods and Software Programs. *The American Journal of Human Genetics 97, 35-53.*

Thornton T, Wu M. 2015. Lecture 7: Introduction to Rare Variant Analysis and Collapsing Tests. *Summer Institute in Statistical Genetics 2015.*

Turner S. 2014. qqman: an R package for creating Q-Q and manhattan plots from GWAS results (https://cran.r-project.org/web/packages/qqman/vignettes/qqman.html)

Voorman A, Brody J, Chen H, Lumley T, Davis B. 2014. seqMeta: an R Package for meta-analyzing region-based tests of rare DNA variants (https://cran.r-project.org/web/packages/seqMeta/index.html)

Wang K. 2016. Boosting the Power of the Sequence Kernel Association Test by Properly Estimating Its Null Distribution. *The American Journal of Human Genetics 99, 104–114.*

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L, and others. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research 2014; 42.*

Wu X, Dong H, Luo L, Zhu Y, Peng G, Reveille JD, Xiong M. 2010. A Novel Statistic for Genome-Wide Interaction Analysis. *PLoS Genet 6(9)*

Xuan J, Yu Y, Quing T, Guo L, Shi L. 2013. Next-generation sequencing in the clinic: Promises and challenges. *Cancer Letters; 340: 284-295.*

# Annex

| Chr | initial | non_monomorphic | rare | info.0.3 | Percentage |
|---|---|---|---|---|---|
| 1 | 18963 | 18694 | 9739 | 9736 | 51.34 |
| 2 | 14335 | 14146 | 6978 | 6978 | 48.68 |
| 3 | 11266 | 11139 | 5237 | 5237 | 46.48 |
| 4 | 8250 | 8122 | 3404 | 3404 | 41.26 |
| 5 | 8831 | 8738 | 3678 | 3678 | 41.65 |
| 6 | 11768 | 11689 | 4788 | 4788 | 40.69 |
| 7 | 9894 | 9790 | 4550 | 4549 | 45.98 |
| 8 | 6856 | 6766 | 2857 | 2857 | 41.67 |
| 9 | 8083 | 7986 | 3996 | 3996 | 49.44 |
| 10 | 8082 | 7960 | 3350 | 3350 | 41.45 |
| 11 | 12137 | 11918 | 6156 | 6155 | 50.71 |
| 12 | 10457 | 10318 | 4892 | 4892 | 46.78 |
| 13 | 3573 | 3517 | 1293 | 1293 | 36.19 |
| 14 | 6848 | 6759 | 3546 | 3546 | 51.78 |
| 15 | 6733 | 6626 | 3455 | 3454 | 51.3 |
| 16 | 8442 | 8312 | 4861 | 4861 | 57.58 |
| 17 | 10417 | 10269 | 6056 | 6055 | 58.13 |
| 18 | 3083 | 3020 | 1032 | 1032 | 33.47 |
| 19 | 13672 | 13574 | 8266 | 8266 | 60.46 |
| 20 | 4770 | 4726 | 2272 | 2272 | 47.63 |
| 21 | 2459 | 2432 | 1111 | 1111 | 45.18 |
| 22 | 4472 | 4441 | 2358 | 2357 | 52.71 |
| **Total** | **193391** | **190942** | **93875** | **93867** | **48.54** |

**Table 1'. Resume of number of SNPs by Chromosome**

| Chr | Burden | SKAT | SKATO | Total Overlap | Overlap Burden-SKAT | Overlap Burden-SKATO | Overlap SKAT-SKATO |
|---|---|---|---|---|---|---|---|
| 1 | 19 | 12 | 18 | 9 | 0 | 6 | 3 |
| 2 | 11 | 10 | 11 | 9 | 0 | 0 | 1 |
| 3 | 4 | 6 | 7 | 2 | 0 | 0 | 4 |
| 4 | 7 | 6 | 7 | 6 | 0 | 1 | 0 |
| 5 | 12 | 8 | 9 | 7 | 0 | 1 | 1 |
| 6 | 11 | 11 | 12 | 8 | 0 | 0 | 3 |
| 7 | 11 | 10 | 10 | 9 | 0 | 0 | 1 |
| 8 | 6 | 6 | 7 | 4 | 0 | 1 | 2 |
| 9 | 3 | 4 | 4 | 3 | 0 | 0 | 1 |
| 10 | 5 | 3 | 4 | 1 | 0 | 1 | 2 |
| 11 | 15 | 16 | 22 | 6 | 1 | 6 | 9 |
| 12 | 11 | 10 | 11 | 8 | 0 | 1 | 2 |
| 13 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | 8 | 6 | 7 | 4 | 0 | 1 | 2 |
| 15 | 10 | 8 | 11 | 7 | 0 | 2 | 1 |
| 16 | 10 | 13 | 12 | 8 | 0 | 0 | 3 |
| 17 | 7 | 5 | 7 | 4 | 0 | 2 | 1 |
| 18 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| 19 | 26 | 26 | 28 | 19 | 1 | 2 | 6 |
| 20 | 1 | 2 | 2 | 1 | 0 | 0 | 1 |
| 21 | 3 | 3 | 3 | 2 | 0 | 0 | 1 |
| 22 | 3 | 3 | 4 | 2 | 0 | 0 | 1 |
| Total | 184 | 169 | 197 | 119 | 2 | 24 | 46 |

**Table 2'. Number of significant genes by method**

| *Symbol* | *pval_Burden* | *pval_SKAT* | *pval_SKATO* | *Description* |
|---|---|---|---|---|
| **CPSF3** | 2.72E-010 | 2.33E-009 | 3.71E-010 | cleavage and polyadenylation specific factor 3 |
| **MFGE8** | 1.23E-009 | 2.31E-009 | 3.08E-010 | milk fat globule-EGF factor 8 protein |
| **KHDRBS1** | 3.61E-010 | 2.20E-008 | 1.06E-009 | KH RNA binding domain containing, signal transduction associated 1 |
| **CTD-2126E3.3** | 2.95E-007 | 5.09E-007 | 2.00E-007 | <NA> |
| **DENND4C** | 7.78E-007 | 3.20E-007 | 4.33E-007 | DENN domain containing 4C |
| **ZNF473** | 3.92E-007 | 6.30E-007 | 5.23E-007 | zinc finger protein 473 |
| **ZBTB8B** | 9.82E-007 | 9.82E-007 | 6.07E-007 | zinc finger and BTB domain containing 8B |
| **TOR3A** | 9.82E-007 | 9.82E-007 | 6.39E-007 | torsin family 3 member A |
| **RP11-328C8.2** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |
| **RP3-335N17.2** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |
| **RP11-371E8.4** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |
| **SOCS5** | 9.82E-007 | 9.82E-007 | 6.39E-007 | suppressor of cytokine signaling 5 |
| **SLC25A32** | 9.82E-007 | 9.82E-007 | 6.39E-007 | solute carrier family 25 member 32 |
| **OR2F1** | 9.82E-007 | 9.82E-007 | 6.39E-007 | olfactory receptor family 2 subfamily F member 1 (gene/pseudogene) |
| **RP11-770J1.4** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |
| **AC005229.7** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |
| **OR10D1P** | 9.82E-007 | 9.82E-007 | 6.39E-007 | olfactory receptor family 10 subfamily D member 1 pseudogene |
| **RP11-290H9.4** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |
| **C6orf47-AS1** | 9.82E-007 | 9.82E-007 | 6.39E-007 | C6orf47 antisense RNA 1 |
| **C6orf47** | 9.82E-007 | 9.82E-007 | 6.39E-007 | chromosome 6 open reading frame 47 |
| **ERC1** | 9.82E-007 | 9.82E-007 | 6.39E-007 | ELKS/RAB6-interacting/CAST family member 1 |
| **AC006994.2** | 9.82E-007 | 9.82E-007 | 6.39E-007 | <NA> |
| **ACADL** | 9.82E-007 | 9.82E-007 | 6.39E-007 | acyl-CoA dehydrogenase, long chain |
| **KCTD17** | 9.82E-007 | 9.82E-007 | 6.39E-007 | potassium channel tetramerization domain containing 17 |
| **CYP27B1** | 1.05E-006 | 1.08E-006 | 5.44E-007 | cytochrome P450 family 27 subfamily B member 1 |
| **BRPF1** | 9.85E-007 | 1.07E-006 | 6.14E-007 | bromodomain and PHD finger containing 1 |
| **CLDN6** | 9.82E-007 | 1.07E-006 | 6.39E-007 | claudin 6 |
| **CRLS1** | 9.82E-007 | 1.07E-006 | 6.39E-007 | cardiolipin synthase 1 |
| **ATP10D** | 9.82E-007 | 1.07E-006 | 6.39E-007 | ATPase phospholipid transporting 10D (putative) |
| **MYPOP** | 9.82E-007 | 1.07E-006 | 6.39E-007 | Myb related transcription factor, partner of profilin |
| **ANKLE1** | 9.82E-007 | 1.07E-006 | 6.39E-007 | ankyrin repeat and LEM domain containing 1 |
| **PMVK** | 9.82E-007 | 1.07E-006 | 6.39E-007 | phosphomevalonate kinase |
| **ARPC1A** | 9.82E-007 | 1.07E-006 | 6.39E-007 | actin related protein 2/3 complex subunit 1A |
| **NPAS4** | 9.82E-007 | 1.07E-006 | 6.39E-007 | neuronal PAS domain protein 4 |
| **CHST6** | 9.82E-007 | 1.07E-006 | 6.39E-007 | carbohydrate sulfotransferase 6 |
| **RP11-77K12.4** | 9.82E-007 | 1.07E-006 | 6.39E-007 | <NA> |
| **TBX21** | 9.82E-007 | 1.07E-006 | 6.39E-007 | T-box 21 |
| **PTGR2** | 9.82E-007 | 1.07E-006 | 6.39E-007 | prostaglandin reductase 2 |
| **ANKRD34B** | 9.83E-007 | 1.07E-006 | 6.39E-007 | ankyrin repeat domain 34B |
| **DPY19L3** | 9.83E-007 | 1.07E-006 | 6.39E-007 | dpy-19 like 3 (C. elegans) |
| **CD3EAP** | 9.82E-007 | 1.07E-006 | 6.40E-007 | CD3e molecule associated protein |
| **DIAPH1** | 9.85E-007 | 1.07E-006 | 6.39E-007 | diaphanous related formin 1 |
| **SLC9A3R1** | 9.85E-007 | 1.07E-006 | 6.39E-007 | SLC9A3 regulator 1 |
| **CALML4** | 1.01E-006 | 1.11E-006 | 6.54E-007 | calmodulin like 4 |
| **RP11-545N8.3** | 1.07E-006 | 1.08E-006 | 6.44E-007 | <NA> |
| **CRYBA2** | 1.13E-006 | 1.13E-006 | 6.79E-007 | crystallin beta A2 |
| **IL32** | 5.89E-007 | 1.95E-006 | 7.14E-007 | interleukin 32 |
| **LAP3P2** | 1.55E-006 | 1.33E-006 | 9.46E-007 | leucine aminopeptidase 3 pseudogene 2 |
| **DBP** | 3.09E-006 | 1.13E-006 | 6.56E-007 | D-box binding PAR bZIP transcription factor |

| *Symbol* | *pval_Burden* | *pval_SKAT* | *pval_SKATO* | *Description* |
|---|---|---|---|---|
| **PRSS30P** | 5.27E-006 | 1.60E-006 | 1.21E-006 | protease, serine, 30 pseudogene |
| **TRIM50** | 1.29E-006 | 4.12E-006 | 2.85E-006 | tripartite motif containing 50 |
| **PPHLN1** | 6.39E-006 | 1.68E-006 | 1.25E-006 | periphilin 1 |
| **LENG1** | 3.38E-006 | 6.13E-006 | 5.09E-007 | leukocyte receptor cluster member 1 |
| **FAM114A1** | 9.44E-006 | 1.73E-006 | 1.11E-006 | family with sequence similarity 114 member A1 |
| **SCYL1** | 1.18E-005 | 1.96E-006 | 1.82E-006 | SCY1 like pseudokinase 1 |
| **NEDD4** | 1.67E-006 | 1.51E-005 | 2.10E-006 | neural precursor cell expressed, E3 ubiquitin protein ligase |
| **CRIP2** | 1.20E-005 | 4.13E-006 | 4.75E-006 | cysteine rich protein 2 |
| **ZNF671** | 6.16E-006 | 1.72E-005 | 1.36E-006 | zinc finger protein 671 |
| **CTC-453G23.5** | 2.32E-005 | 3.87E-006 | 4.84E-006 | <NA> |
| **TRIM15** | 9.83E-006 | 1.87E-005 | 6.88E-006 | tripartite motif containing 15 |
| **C2orf40** | 3.47E-005 | 1.15E-006 | 5.82E-007 | chromosome 2 open reading frame 40 |
| **ANXA3** | 1.53E-005 | 1.98E-005 | 2.07E-005 | annexin A3 |
| **PAK1** | 6.40E-008 | 9.18E-005 | 3.22E-006 | p21 (RAC1) activated kinase 1 |
| **CEP89** | 6.50E-005 | 3.09E-005 | 3.70E-005 | centrosomal protein 89 |
| **PLAUR** | 9.68E-006 | 5.39E-005 | 7.06E-005 | plasminogen activator, urokinase receptor |
| **SLC44A4** | 5.65E-006 | 1.16E-004 | 6.42E-005 | solute carrier family 44 member 4 |
| **SLC22A5** | 1.15E-004 | 7.14E-005 | 2.13E-006 | solute carrier family 22 member 5 |
| **HIGD1C** | 6.36E-005 | 6.36E-005 | 6.33E-005 | HIG1 hypoxia inducible domain family member 1C |
| **MRPL1** | 6.36E-005 | 6.36E-005 | 6.33E-005 | mitochondrial ribosomal protein L1 |
| **UGT3A2** | 6.36E-005 | 6.36E-005 | 6.33E-005 | UDP glycosyltransferase family 3 member A2 |
| **C19orf73** | 6.36E-005 | 6.36E-005 | 6.33E-005 | chromosome 19 open reading frame 73 |
| **CTD-2353F22.1** | 6.36E-005 | 6.36E-005 | 6.33E-005 | <NA> |
| **SLC1A3** | 6.36E-005 | 6.36E-005 | 6.33E-005 | solute carrier family 1 member 3 |
| **RP11-529J17.3** | 6.36E-005 | 6.36E-005 | 6.33E-005 | <NA> |
| **RP11-312J18.5** | 6.36E-005 | 6.36E-005 | 6.33E-005 | <NA> |
| **RP11-80A15.1** | 6.36E-005 | 6.36E-005 | 6.33E-005 | <NA> |
| **MPZ** | 6.36E-005 | 6.36E-005 | 6.33E-005 | myelin protein zero |
| **U2AF1L4** | 6.36E-005 | 6.36E-005 | 6.35E-005 | U2 small nuclear RNA auxiliary factor 1 like 4 |
| **RP5-1063M23.1** | 6.36E-005 | 6.36E-005 | 6.35E-005 | <NA> |
| **TMEM106B** | 6.36E-005 | 6.36E-005 | 6.35E-005 | transmembrane protein 106B |
| **RP11-350O14.18** | 6.36E-005 | 6.36E-005 | 6.35E-005 | <NA> |
| **TMEM210** | 6.36E-005 | 6.36E-005 | 6.35E-005 | transmembrane protein 210 |
| **CADM4** | 6.39E-005 | 6.39E-005 | 6.39E-005 | cell adhesion molecule 4 |
| **KRT31** | 6.36E-005 | 6.86E-005 | 6.33E-005 | keratin 31 |
| **NAPEPLD** | 6.36E-005 | 6.86E-005 | 6.33E-005 | N-acyl phosphatidylethanolamine phospholipase D |
| **HMGN1** | 6.36E-005 | 6.86E-005 | 6.35E-005 | high mobility group nucleosome binding domain 1 |
| **PGBD2** | 6.36E-005 | 6.86E-005 | 6.35E-005 | piggyBac transposable element derived 2 |
| **CRCP** | 6.36E-005 | 6.86E-005 | 6.35E-005 | CGRP receptor component |
| **RP5-1132H15.1** | 6.36E-005 | 6.86E-005 | 6.35E-005 | <NA> |
| **RASGRP3** | 6.45E-005 | 6.86E-005 | 6.36E-005 | RAS guanyl releasing protein 3 |
| **SCGB2B2** | 6.59E-005 | 9.76E-005 | 3.46E-005 | secretoglobin family 2B member 2 |
| **FAM131C** | 6.57E-005 | 6.91E-005 | 6.45E-005 | family with sequence similarity 131 member C |
| **MCTP1** | 7.19E-005 | 6.87E-005 | 6.36E-005 | multiple C2 and transmembrane domain containing 1 |
| **SIK1** | 8.18E-005 | 6.02E-005 | 6.25E-005 | salt inducible kinase 1 |
| **ZFYVE19** | 7.26E-005 | 6.92E-005 | 6.45E-005 | zinc finger FYVE-type containing 19 |
| **AC106782.20** | 1.08E-004 | 4.52E-005 | 5.70E-005 | <NA> |
| **BCL7C** | 1.08E-004 | 4.52E-005 | 5.70E-005 | BCL tumor suppressor 7C |
| **MIR4519** | 1.08E-004 | 4.52E-005 | 5.70E-005 | microRNA 4519 |

| *Symbol* | *pval_Burden* | *pval_SKAT* | *pval_SKATO* | *Description* |
|---|---|---|---|---|
| **FAM217A** | 6.26E-005 | 1.22E-004 | 5.94E-005 | family with sequence similarity 217 member A |
| **PI15** | 1.03E-004 | 8.63E-005 | 8.14E-005 | peptidase inhibitor 15 |
| **RP11-758M4.4** | 1.03E-004 | 8.63E-005 | 8.14E-005 | <NA> |
| **OR2T8** | 2.21E-007 | 2.84E-004 | 6.88E-006 | olfactory receptor family 2 subfamily T member 8 |
| **RP1-111C20.4** | 2.21E-007 | 2.84E-004 | 6.88E-006 | <NA> |
| **ZWILCH** | 2.90E-004 | 4.03E-006 | 4.77E-006 | zwilch kinetochore protein |
| **ERCC1** | 9.28E-005 | 2.06E-004 | 7.78E-005 | ERCC excision repair 1, endonuclease non-catalytic subunit |
| **LRIT2** | 9.28E-005 | 2.06E-004 | 7.80E-005 | leucine rich repeat, Ig-like and transmembrane domains 2 |
| **ZNF589** | 6.24E-005 | 2.62E-004 | 6.17E-005 | zinc finger protein 589 |
| **ASB9P1** | 8.40E-006 | 3.88E-004 | 2.33E-005 | ankyrin repeat and SOCS box containing 9 pseudogene 1 |
| **TMEM120B** | 1.09E-004 | 2.22E-004 | 9.39E-005 | transmembrane protein 120B |
| **LIG1** | 3.37E-004 | 5.46E-005 | 7.44E-005 | DNA ligase 1 |
| **HOXA2** | 7.75E-005 | 1.89E-004 | 2.20E-004 | homeobox A2 |
| **C22orf43** | 3.59E-004 | 1.18E-004 | 1.08E-004 | <NA> |
| **AC226119.4** | 5.43E-004 | 6.74E-005 | 6.24E-005 | <NA> |
| **AC226119.5** | 5.43E-004 | 6.74E-005 | 6.24E-005 | <NA> |
| **HOXD9** | 3.86E-005 | 4.11E-004 | 4.30E-004 | homeobox D9 |
| **IGKV1-8** | 2.70E-004 | 3.09E-004 | 3.66E-004 | immunoglobulin kappa variable 1-8 |
| **NFKBID** | 1.09E-004 | 5.16E-004 | 5.98E-004 | NFKB inhibitor delta |
| **RP11-101E19.8** | 4.31E-004 | 4.05E-004 | 4.24E-004 | <NA> |
| **AAMDC** | 2.00E-004 | 5.51E-004 | 6.02E-004 | adipogenesis associated Mth938 domain containing |

**Table 3'. *P-value* and description of 119 significant genes**