SNP association in coronary disease

Neus Amat Sorando, Aleix Canalda Baltrons, Maria Díaz Ros, Judit Tella Vila

Contents

1 Abstract

The coronary heart disease is developed when the major blood vessels that supply the heart become damaged. The objective of this analysis is to identify SNPs associated with this disease. To do so, a Genome Wide Association Study was performed and six significant SNPs were found inside a gene region. Out of these, there is one gene (*PDE4D*) that seems to be more directly related to the disease.

2 Introduction

Coronary heart disease or coronary artery disease (**CAD**) is the most common cardiovascular problem, being the leading cause of death in USA (without any difference between sexes).

This disease consists on the hardening and narrowing of arteries. The major causal of this event is the plaques of cholesterol in the inner walls. These plaques could have other waste components. This buildup is called atherosclerosis.

As the accumulation of cholesterol grows, less blood can flow thought the arteries causing lack of oxygen where is needed. The CAD could lead to arrhythmia and hearth attack.

Coronary artery disease has a complex etiology and usually has risk factors and genetic predisposition (with a predisposition inheritance between 40 and 50%).

The aim of this analysis is to identify SNPs correlated with a higher predisposition in CAD.

3 Methods

To identify SNPs that are linked with a higher risk of CAD we have done a **GWAS** (Genome Wide Association Study).

3.1 Packages and tools used

For the analysis of our data we have used the following `packages` on R and/or R studio:

- **ggplot2**. It's a system to create graphs. You have to provide the data and specify how to map variables to aesthetics.
- **dplyr**. It's a grammar of data manipulation, providing a set of verbs that help to solve the most common data manipulation challenges, for example select or filter some data.
- **ggrepel**. Provides text and label geoms for 'ggplot2' that help to avoid overlapping text labels.
- **devtools**. It's a collection of package development tools aiming to simplify their development providing R functions that simplify common tasks.

- **SNPassoc**. It carries out most common analysis when performing whole genome assotiation studies. These analyses include descriptive statistics and exploratory analysis of missing values, calculation of Hardy-Weinberg equilibrium and analysis of assotiation based on generalized linear models.
- **BiocManager**. It allows users to install and mange packages from the Bioconductor project. Bioconductor focuses on the statistical analysis and comprehension of high-throughput genomic data.
- **snpStats**. Contains classes and statistical methods for large SNP assotiation studies. It also allows for uncertainty in genotypes.
- **SNPRelate**. It provides a binary format for SNP data in GWAS. I's also designed to accelerate two key computations on SNP data using parallel computing.

3.2 Data description

We start off with three main fields of information on the individuals we're analyzing, which would be "genotypes", "fam" and "map". Inside "genotypes" we can find that we have 425 individuals (number of rows) and 582892 SNPs (number of columns). Inside "fam" we can see the information related to an individual's family. Finally, in "map", we find information on each SNP. Next, we have to merge the phenotypic information on the individuals with their respective genotypic information, but we have to see if the individuals in both datasets match. If they don't match, we'll only keep those individuals that actually have both types of information. In our case, the individuals didn't match, so we had to take these previously mentioned steps.

3.3 Quality Control

In order to follow through with the GWAS analysis, we must do a quality control of both the SNPs and the individuals and filter those that don't meet certain thresholds or requirements, in order to achieve a more accurate analysis.

3.3.1 SNPs

To perform a quality control of the SNPs we're going to get rid of them using three different criteria:

- SNPs with high rate of missing, typically markers with a call rate smaller than 95 percent.
- SNPs with a minor allele frequency less than 5 percent.
- SNPs that do not pass the Hardy-Weinberg equilibrium test.

After applying some coding lines to filter as mentioned above, 4443 SNPs were removed for a bad call rate, 65183 SNPs were removed due to a low MAF and 234 SNPs did not pass the Hardy-Weinberg test, making a total of 68.998 SNPs removed out of 100.000 after applying the filter.

3.3.2 Individuals

When doing a quality control on the individuals we're going to be using 4 different criteria:

- Identification of individuals with discordant reported and genomic sex.
- Identification of individuals with outlying missing genotype or heterozygosity.
- Identification of duplicated or related individuals.
- Identification of individuals of divergent ancestry from the sample.

First we start by removing the individuals with a discordant reported and genomic sex. This is usually done by inferring heterozygosity of chromosome X, where males are expected to present 0 heterozygosity and females are expected an approximate value of 0,3. We can see the result of this filter in Figure 1, where we can observe that three individuals who were reported as male present a heterozygosity of 0,3, therefore are chromosomally females.

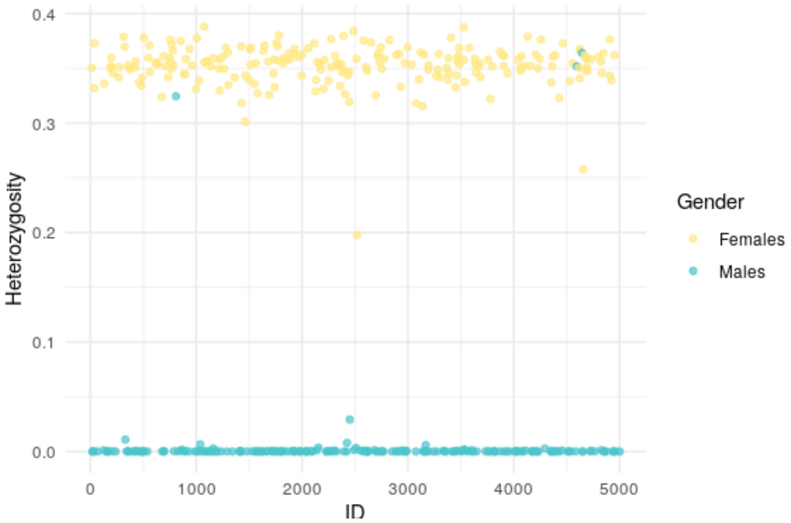

Figure 1 – Heterozygosity of chromosome X of each individual present in our dataset.

The next step is to filter those individuals with missing genotypes and also with an overall outlying heterozygosity score. This is done on the one hand by filtering those individuals with NA in more than 5 percent of their genotypes, and on the other hand by creating the statistic value F, which is done by applying: 1 - f(Aa)/Exp[f(Aa)]. Those that present an F value of ±0.1, means that their heterozygosity rate is lower than 0.32 and will be filtered. In Figure 2 we can observe that in our sample we have two individuals that present more than 0.1 in their F value, and will therefore be filtered.

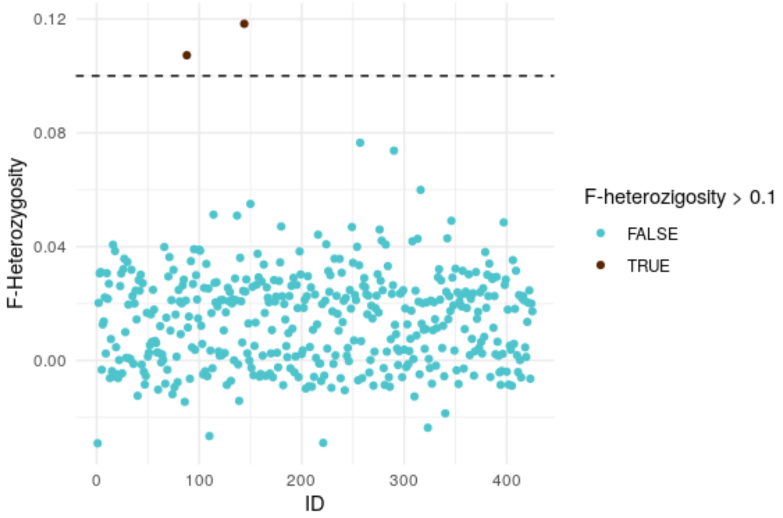

Figure 2 – F value for each individual in our data set.

Finally, we want to also filter those individuals that are related to each other since a GWAS is a population study with random samples, and related individuals don't contribute correctly. By applying a threshold of 0.1 on kinship we filtered two individuals.

After this process, we ended up filtering 2 individuals with >5 percent missing genotypes, 3 for sex discrepancies, 2 for outlying heterozygosity and 2 for having >0.1 kinship.

4 Results

After having filtered all the data we were able to run a GWAS test using a quantitative trait (age, specifically).

A useful way to visualize the GWAS results would be by obtaining a Manhattan plot. Before we can create it we need to prepare the data for such a plot. For example, for the x axis we want to order the SNPs in a way that the first SNP is the first one found in chromosome 1 and the last SNP is the last found in chromosome 22. That can be done with an R loop and creating a single column in our data set.

We also want to apply the Bonferroni correction to our data. Now we're ready to create the plot and design it to our liking.

After doing all the analyses we obtained this plot:

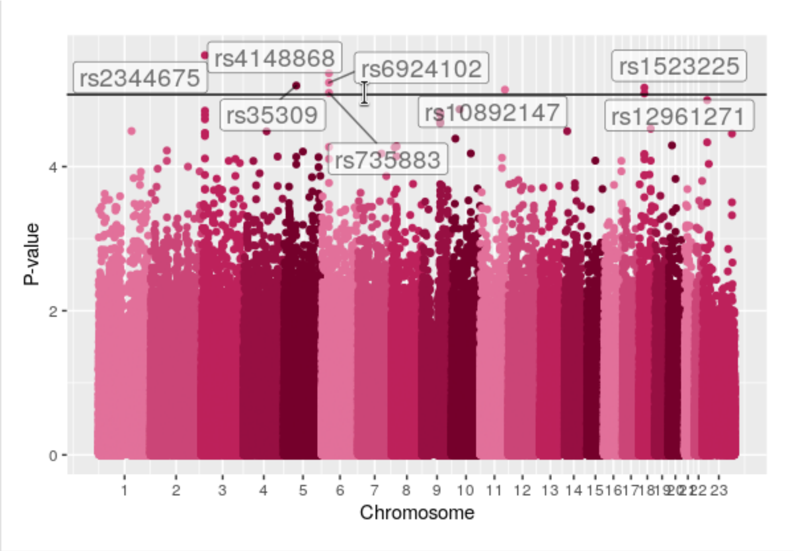

Figure 3 – Manhattan plot of the genome-wide association studies (GWAS) for coronary diseases

We can see that there are eight significant SNPs chromosomes 3, 5, 6, 11 and 18, the most significant being in the chromosome 3.

Once we had this results, we visualized the position of every SNP in the genome using LocusZoom. This way we can see if the SNPs are located inside a gene or not.

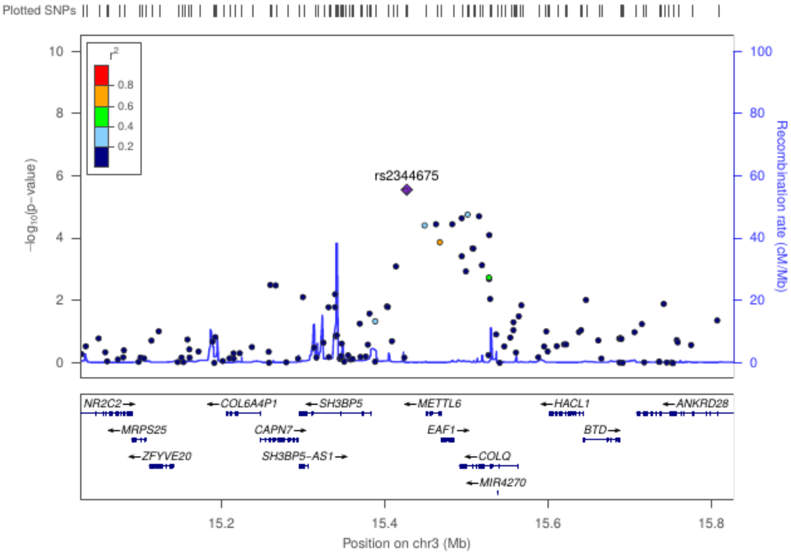

Figure 4 – Location of the rs2344675 SNP

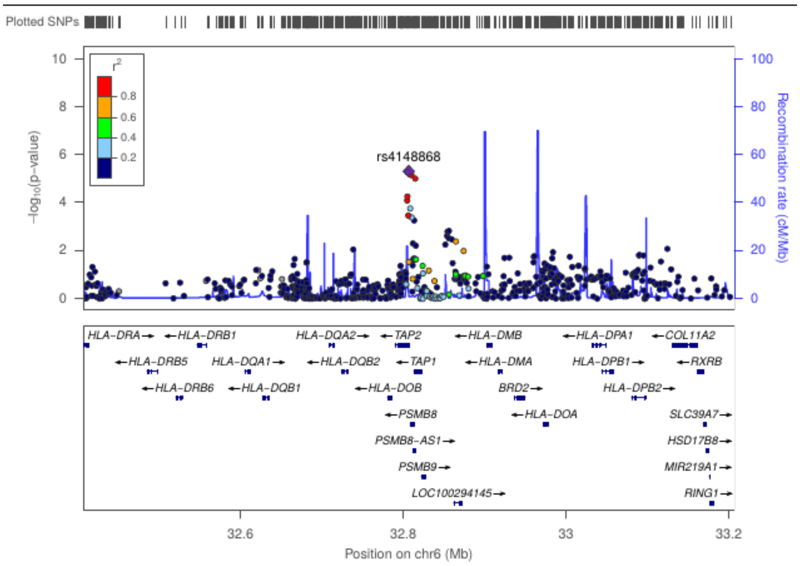

Figure 5 – Location of the rs4148868 SNP

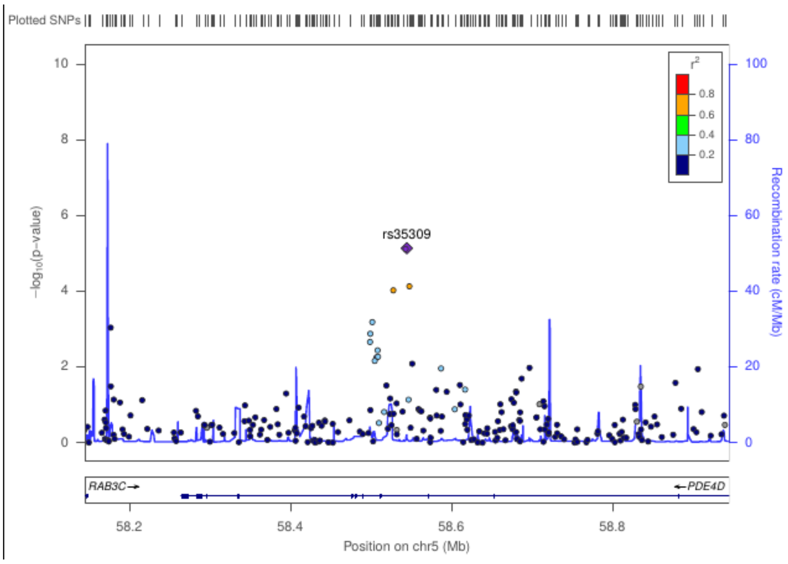

Figure 6 – Location of the rs35309 SNP

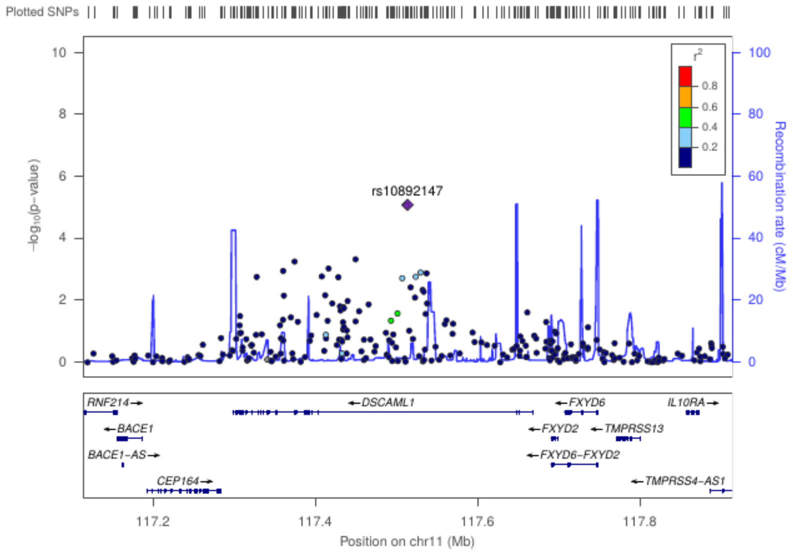

Figure 7 – Location of the rs10892147 SNP

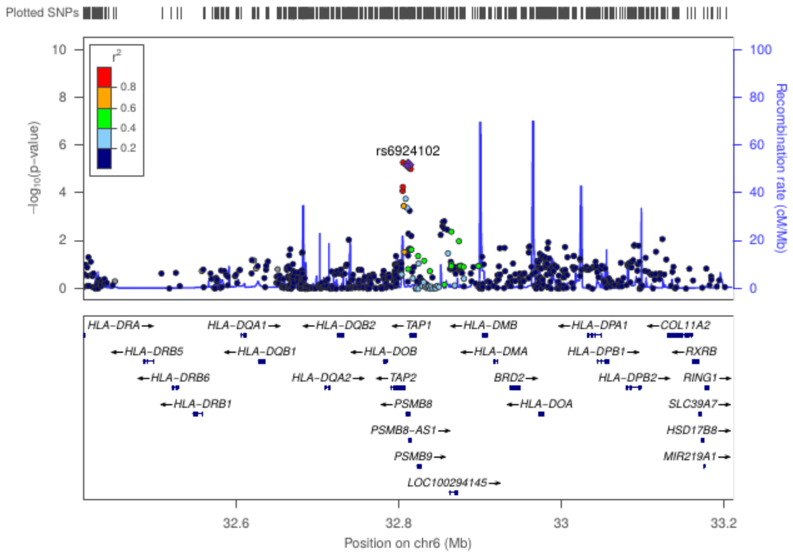

Figure 8 – Location of the rs6924102 SNP

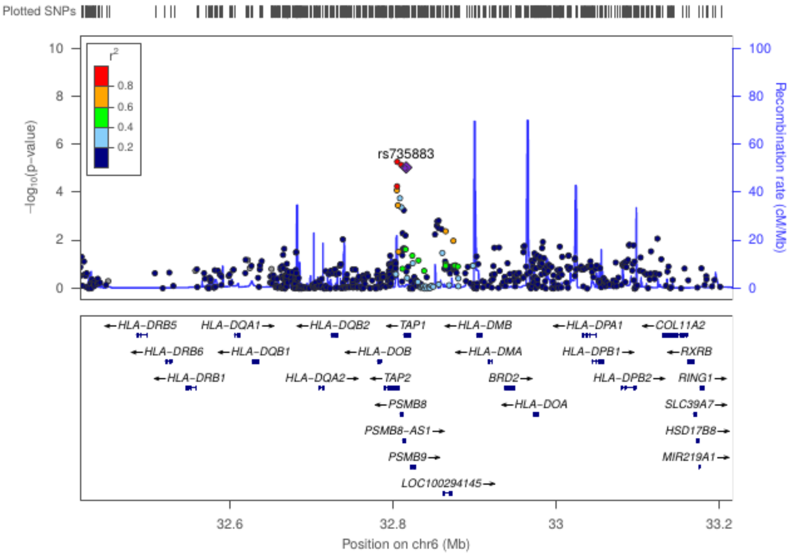

Figure 9 – Location of the rs735883 SNP

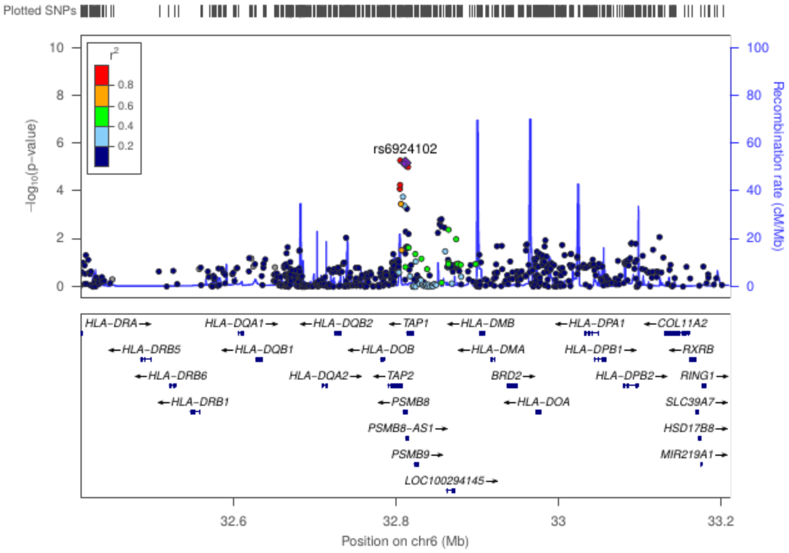

Figure 10 – Location of the rs6924102 SNP

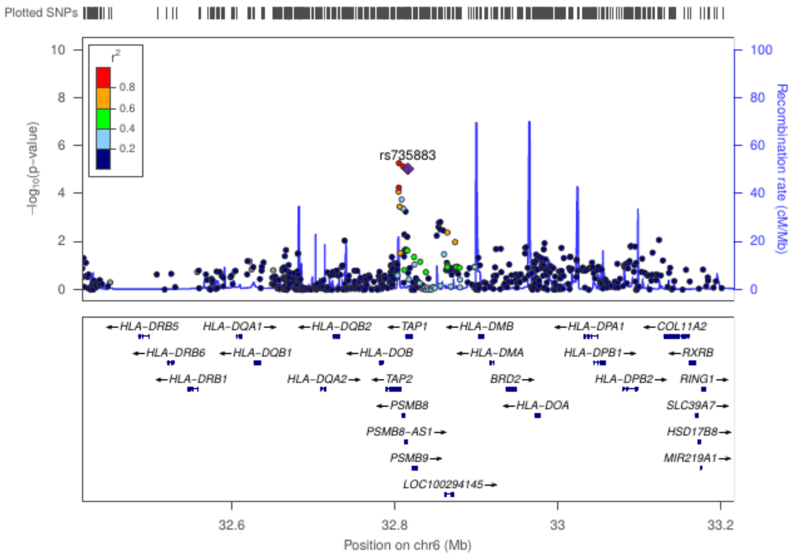

Figure 11 – Location of the rs735883 SNP

As we can see in the pictures, six from the eight significant SNPs are located inside a gene.

5 Discussion

The significant SNPs were rs2344675, rs35309, rs4148868, rs6924102, rs735883, rs10892147, rs1523225 and rs12961272. The next step will consist on searching the SNPs in NCBI.

The first SNP (**rs2344675**), which is located in chromosome 3 (chr3:15385610), it's a transition (G>A) in an intronic region of *METTL6* gene. The allele which has a guanine has a relatively high frequency, around 0.2 and doesn't have any clinical significance.

The following SNP, **rs35309**, is located in chromosome 5 (chr5:59248822) and it's a transition (T>C) or a transversion (T>A) in an intronic region of *PDE4D* gene. The frequency of the alleles is different between populations comprising relatively high (0.222) and low (0.06) values. However, it hasn't been reported any clinical significance.

The next SNP, **rs4148868**, is located in chromosome 6 (chr6:32838807) and it's a transition (G>A) in the 5' UTR of *TAP2* gene. The frequency of the adenine allele is around 0.45, a high frequency. There isn't any report of clinical significance and besides, the allele frequencies are too high to be related to any disease.

Another significant SNP, **rs6924102**, is located in chromosome 6 (chr6:32843606) and it's a transition (A>G) in an intronic region of *PSMB8* gene in the minus strand and an upstream variant in the plus strand (*PSMB8-AS1*, antisense RNA1). The guanine allele frequency is around 0.43, being too high to have a clinical effect. Furthermore, there isn't any clinical significance reported.

The SNP **rs735883**, which is located in chromosome 6 (chr6:32848277) as well, consists in a transition (G>A) on an intronic region of *TAP1* gene. The allele frequency is high, having values comprised between 0.356 and 0.48 for the adenine allele. It hasn't been reported as a clinical significant SNP.

The following SNP, **rs10892147**, is located in chromosome 11 (chr11:117643086) and it's a transition (T>C) in an intronic region of *DSCAML1* gene. The cytosine allele frequency is around 0.16 (except in the case of Vietnamese where it has a frequency of 0.07). This variation hasn't been reported as a clinical significant variant.

Other significant SNP that we detected was **rs1523225** and it's located in chromosome 8 (chr18:28323839) as a transition (T>C) without any reported consequence in any gene. The cytosine variant comprises frequencies between 0.13 and 0.327 in different populations. As expected, there isn't any reported clinical significance.

The last identified SNP, **rs12961272**, is located in chromosome 18 (chr18:77743149) and it's a transition (A>G) without any reported consequence in any gene. The allele's frequency between populations differs a lot, comprising values between 0.81 and 0.281 for the adenine allele. Because there isn't any gene consequence reported, neither are clinical effects.

Having in mind that none of the SNPs that we were able to identify as significant to coronary disease haven't been reported in ClinVar (database of reports that links human variation and health status) we will search the mentioned genes. Maybe one SNP alone couldn't trigger the disease, but the combination of some of them could:

- *METTL6* (methyltransferase like 6): it's a coding gene which transcript codifies for a protein with tRNA (cytosine-5-)-methyltransferase activity related in tRNA C5-cytosine methylation.

  The gene expression is higher in testis and brain.

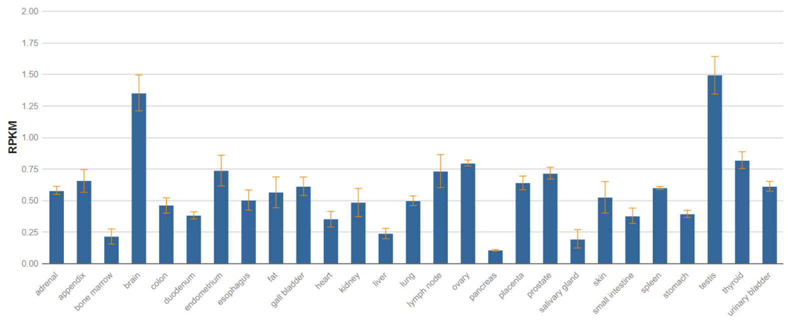

Figure 12 – Differential *METTL6* expression between tissues.

- *PDE4D* (phosphodiesterase 4D): is a protein coding gene with 3',5'-cyclic-AMP phosphodiesterase activity, ATPase binding, beta-2 adrenergic receptor binding, cAMP binding, drug binding, enzyme binding, ion channel binding, metal ion binding, protein binding and scaffold protein binding.

  This gene is involved in the following processes: G protein-coupled receptor signaling pathway, T cell receptor signaling pathway, adenylate cyclase-activating adrenergic receptor signaling pathway involved in positive regulation of **heart rate**, **adrenergic** receptor signaling pathway, aging, cAMP catabolic process, cAMP-mediated signalling, cellular response to **epinephrine** stimulus, cellular response to lipopolysaccharide, establishment of endothelial barrier, multicellular organism growth, negative regulation of cAMP-mediated signalling, negative regulation of **heart contraction**, negative regulation of peptidyl-serine phosphorylation, negative regulation of **relaxation of cardiac muscle**, neutrophil chemotaxis, positive regulation of interferon-gamma production, positive regulation of interleukin-2 production, positive regulation of interleukin-5 production, regulation of **cardiac muscle cell contraction**, regulation of **cell communication by electrical coupling involved in cardiac conduction and regulation of heart rate**.

The gene expression is higher in bone marrow which has the suport moving function, blood cells production and fat storage.

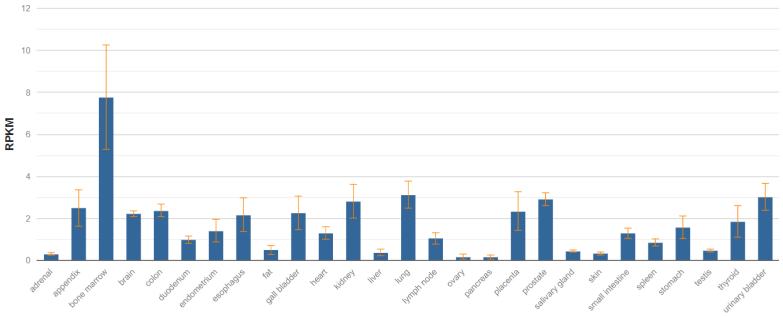

Figure 13 – Differential *PDE4D* expression between tissues.

- *TAP1* (transporter 1, ATP binding cassette subfamily B member): is a membrane-associated protein and it is a member of the superfamily of ATP-binding cassette (ABC) transporters.

  This transporters move various molecules between the inside and the outside of the membranes.This protein is a member of the MDR/TAP subfamily which are involved in multidrug resistance, specifically on the pumping of degraded peptides across the endoplasmic reticulum into themembrane-bound compartment

  The functions of *TAP1* gene are the ADP binding, ATP binding, ATPase activity, ATPase activity coupled to transmembrane movement of substances, MHC class I protein binding, MHC class Ib protein binding, TAP1 binding, TAP2 binding, peptide antigen binding, peptide antigen-transporting ATPase activity, peptide transmembrane transporter activity, peptide-transporting ATPase activity, protein binding and protein homodimerization activity.

  The processes which is involved are the following ones: adaptive immune response, antigen processing and presentation of endogenous peptide antigen via MHC class I, antigen processing and presentation of exogenous peptide antigen via MHC class I TAP-dependent, antigen processing and presentation of peptide antigen via MHC class I, cytosol to ER transport, defense response, peptide transport, protein transport, transmembrane transport, vesicle fusion with endoplasmic reticulum-Golgi intermediate compartment (ERGIC) membrane and in viral process.

  The expression of this gene is higher in spleen, apendix and lymph node.

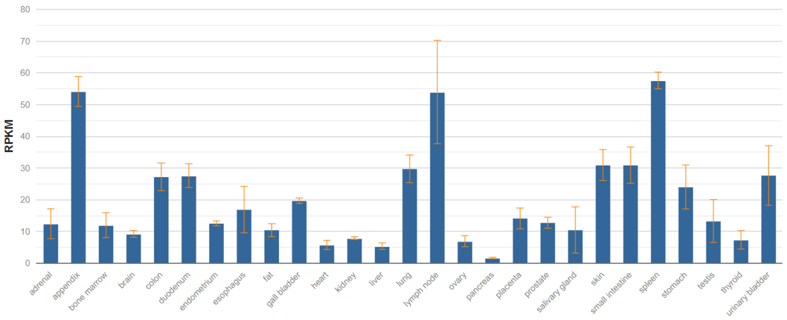

Figure 14 – Differential *TAP1* expression between tissues.

- *TAP2* (transporter 2, ATP binding cassette subfamily B member): is a membrane-associated protein and it is a member of the superfamily of ATP-binding cassette (ABC) transporters.

  As mentioned before, this transporters move various molecules between the inside and the outside of the membranes. This protein is a member of the MDR/TAP subfamily, specifically on the antigen presentation and movement of peptides between cytoplasm and endoplasmic reticulum.

  The transporter is associated to the following processes: ATP binding, ATPase activity, ATPase activity coupled to transmembrane movement of substances, MHC class Ib protein binding, TAP1 binding, peptide antigen-transporting ATPase activity, contributes to peptide transmembrane transporter activity, peptide-transporting ATPase activity, protein binding and tapasin binding.

  The expression of TAP2 gene is higher in appendix, lymph node and spleen.

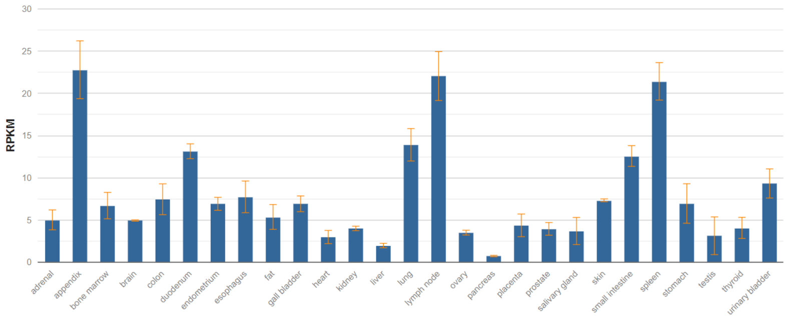

Figure 15 – Differential *TAP2* expression between tissues.

- *PSMB8* (proteasome 20S subunit beta 8): is a multicatalytic proteinase complex with a highly ordered ring-shaped 20S core structure which funtions are endopeptidase activity, protein binding and threonine-type endopeptidase activity.

  The process which is involved the protein are Fc-epsilon receptor signaling pathway, MAPK cascade, NIK/NF-kappaB signalling, SCF-dependent proteasomal ubiquitin-dependent protein catabolic process, T cell receptor signaling pathway, Wnt signaling pathway, planar cell polarity pathway, anaphase-promoting complex-dependent catabolic process, antigen processing and presentation of exogenous peptide antigen via MHC class I TAP-dependent, fat cell differentiation, interleukin-1-mediated signaling pathway, negative regulation of G2/M transition of mitotic cell cycle, negative and positive regulation of canonical Wnt signaling pathway, post-translational protein modification, proteasomal protein catabolic process, proteasomal ubiquitin-independent protein catabolic process, proteasome-mediated ubiquitin-dependent protein catabolic process, protein deubiquitination, protein polyubiquitination, regulation of cellular amino acid metabolic process, regulation of endopeptidase activity, regulation of hematopoietic stem cell differentiation, regulation of mRNA stability and regulation of mitotic cell cycle phase transition.

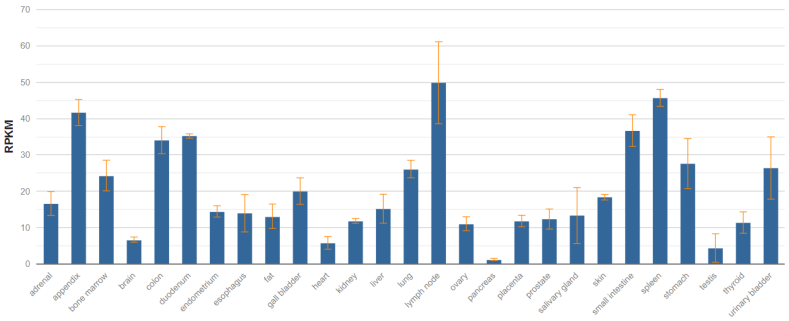

Figure 16 – Differential *PSMB8* expression between tissues.

- *DSCAML1* (DS cell adhesion molecule like 1): is a protein coding gene which forms part of Ig superfamily of cell adhesion molecules and is involved in neuronal differentiation.

  The functions are cell-cell adhesion mediator activity and protein homodimerization activity.

  The molecule is involved in process as axon guidance, axonogenesis, brain development, cell fate determination, central nervous system development, dendrite self-avoidance, dorsal/ventral pattern formation, embryonic skeletal system morphogenesis and homophilic cell adhesion via plasma membrane adhesion molecules.

  The expression of this protein is located especially in brain, ovary and kidney.

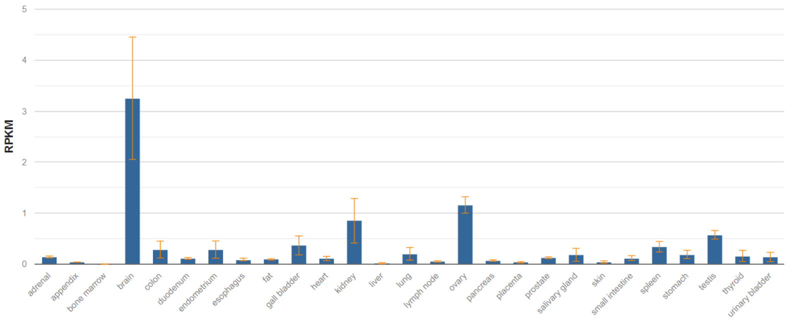

Figure 17 – Differential *DSCAML1* expression between tissues.

With the available information we only confirm one possible SNP that may cause an increse risk of coronary disease, the SNP located in *PDE4D* gene (rs35309) because it's the unique that affects a gene involved with hearth health.

It would be needed more studies with more data to confirm if the identified SNP is also significant in other population or patients and finding more SNPs that we weren't able to identify (we must keep in mind that this disease is complex and probably the presence or absence of the illness is due to the effect of several SNPs).