Which are the variants behind colorectal cancer?

Karmele Alapont, Marina Bataller, Nerea Carrón and Judit Garcia

January 8, 2020

1 Abstract

Complex diseases, like cancer, can be caused by many different genetic variations. This fact makes it very difficult to study and characterize correctly them. A good method to do it is GWAS, a type of analysis that is able to identify genetic variants that cause susceptibility for a disease. We performed this analysis with colorectal cancer data, and found 11 genetic variants that are potentially associated with this disease. We performed a bibliographic research for them, trying to learn more about them and decide which ones would be more interesting to start with further molecular studies.

2 Introduction

GWAS, or Genome-Wide Association Studies, are hypothesis-free methods used in genetics research to associate specific genetic variations with traits, including susceptibility for particular diseases. This method involves scanning genomes of a large number of subjects: people who have the disease and people who don't. You search the whole genome for single nucleotide polymorphisms (SNPs), and you try to find the ones where there's a consistent presence in the subjects with the disease than in people without it. This way, GWAS can identify genetic markers that can predict the presence of this trait or disease.

The technique requires a number of statistical tests that help identify the genetic variations responsible of the disease. Due to its statistical nature, GWAS requires many subjects to have reliable enough results. With the completion of the Human Genome Project and the International HapMap Project, and the development of next-generation sequencing technologies, researchers now have enough data to make this method a good way of finding new loci related to complex diseases.

We performed a GWAS analysis using data from a case-control study. We tried finding genetic variations associated with colorectal cancer, a type of complex disease perfect for this type of analysis.

3 Methods

3.1 Packages and tools used

For this study we used the Rstudio integrated development environment (IDE) for R. We used different R packages to manage our data, analyze it and create graphics to visualize it:

- ggplot2: data visualization package, used to create graphics.
- dplyr: data manipulation package designed to abstract over how the data is stored.
- ggrepel: provides geoms for ggplot2 to repel overlapping text labels.
- devtools: collection of package development tools.
- isglobal-brge: plots GWAS models to ggplot2.
- SNPassoc: genetic association studies. Contains classes and methods to help the analysis of whole genome association studies.
- BiocManager: tool to install and update Bioconductor packages.
- snpStats: analyses genetic-epidemiology studies of association using SNPs.
- SNPRelate: provides a binary format for SNP data in GWAS.

3.2 Data Description

We downloaded the PLINK colorectal cancer data and obtained three different documents from here:

- Colorectal.bed → contains the genomic SNP data
- Colorectal.bim → contains the SNP annotation
- Colorectal.fam → contains the individual's family information

We uploaded this data to R and grouped it in a dataset called colorectal.plink. Then we separated each information group into:

- Colorectal.genotype → genomic SNP data
- Individuals → family information

- Annotation → SNP annotation

To have the full phenotype information we downloaded additional text files and uploaded to R with the name colorectal.phenotype.

In order to analyse the genotypes, SNP annotation, individual's family information and phenotype, we needed to make sure that we had the same individuals and in the same order at the three datasets, so we compared the Ids at the three groups of data. Using the following command:

```
#We check if the rownames of the two objects are identical
    identical(rownames(colorectal.phenotype),
    rownames(colorectal.genotype))
```

Even if the result of this was that the Ids were the same, just in case, we performed the filtering and eliminated the ones that were not interesting. Once we did this, we sorted the ids into the same order. We used the script:

```
ids<-intersect(rownames(colorectal.phenotype),rownames(colorectal.genotype))
genotype <- colorectal.genotype[ids, ]
phenotype <- colorectal.phenotype[ids, ]
identical(rownames(phenotype), rownames(genotype))
individuals <- individuals[ids, ]
```

So now, we only have the data we want to analyse and it is sorted by Id.

3.3 Quality control

Before starting with the GWAS analysis we had to perform some quality controls to make sure that the data we had was good enough for this analysis.

We did this analysis at two different levels:
1. Quality control of SNPs
2. Quality control of individuals

Quality control of SNPs:

Before starting with the analysis, we need to have the control individuals, these are subjects that do not have colorectal cancer (cascon==0) and that have data at this column. So we created the controls dataset with the information from these individuals.

But now we want to have only the genotype information of these individuals. We stored this at genotype.control. Once we separated the control, we pass the data through different measures to perform the quality control:

- SNPs with high rate of missing → SNPs with a call rate less than 95% are removed
- Rare SNPs (MAF) → SNPs with less than 5% minor allele frequency (MAF) are delete
- SNPs that do not pass the Hardy-Weinberg equilibrium test → controls with a Z-value bigger than 3.3 are also removed from the analysis.

In order to do this, we used the following script:

```
use <- info.snps$Call.rate > 0.95 &
   info.snps$MAF > 0.05 &
   abs(info.controls$z. < 3.3)
   mask.snps <- use & !is.na(use)
```

So now we only keep the ones that pass the QC at genotype.qc.snps

To really know how many data you are deleting from your dataset, it is interesting to count the number of individuals that you are not maintaining, we did this using the following script:

```
# Number of SNPs removed for a bad call rate
sum(info.snps$Call.rate < 0.95, na.rm = TRUE)
# Number of SNPs removed for low MAF
sum(info.snps$MAF < 0.05, na.rm = TRUE)
# Number of SNPs removed that do not pass HWE
sum(abs(info.controls$z.HWE > 3.3), na.rm = TRUE)
# The total number of SNPs removed for any reason
sum(!mask.snps)
```

Which means that we only count them if the condition mentioned in the line is true. Finally, we eliminated a sum of 11479 SNPs and kept 88521 from 100,000 SNPs.

Quality control of individuals:

Now we will look at the information at individuals' level. Here we followed four steps.

First, we removed individuals with **sex discrepancies**. Usually, gender is inferred from the heterozygosity of chromosome X, males have an expected heterozygosity of 0 and females of 0.30. So, using 'row.summary' we extracted the heterozygosity.

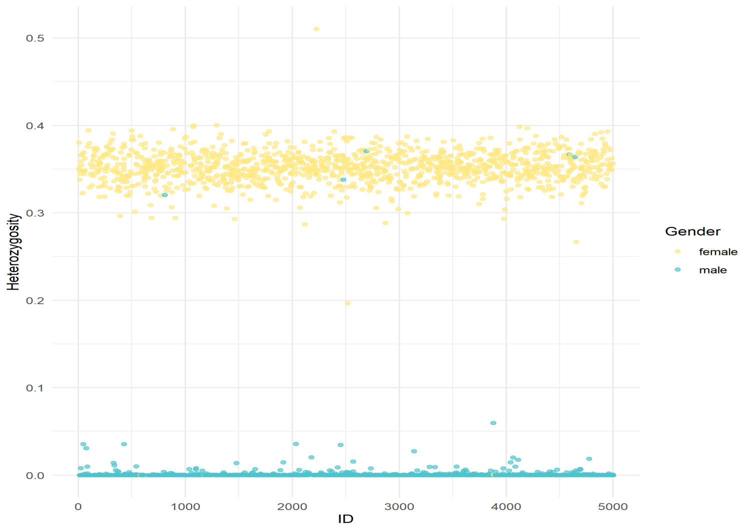

Figure 1: Heterozygosity of chromosome X

Looking at the plot (Figure 1) we can see how there are some males that have higher heterozygosity than expected and females with lower. We located them at sex.discrep to then remove them from the sample.

Once we have this, we identify individuals with **outlying heterozygosity** from the overall genomic heterozygosity. Individuals having a heterozygosity rate lower than 0.32 are considered sample outliers and we want to remove them from the sample.

The next step is to delete **close familial relatedness** between individuals, because it won't be representative of the sample. In order to do this, we used the R package 'SNPRelate' which computes kinship within the sample. As this package requires data in GDS format, we obtained this using the 'snpgdsBED2GDS' function. Then, to remove the related ones, we used the 'snpdgLDpruning' function, this removes iteratively adjacent SNPs that exceed an LD threshold.

The individuals who are candidates to be removed because of relatedness are the ones that have an expected relatedness, which are the ones that have a kinship score > 0.1. To remove the ids of the related individuals we used the function called 'related'.

Then we eliminated individuals with **high rate of missing**. Meaning, individuals with a call rate less than 95

So, what we did was to remove individuals with more than 5% missing genotypes, with sex discrepancies, F-heterozigosity absolute value >0.1 and kindship coefficient >0.1 from the genotype and phenotype data. We did this using the following script:

```
use<- info.indv$Call.rate > 0.95 &
abs(info.indv$hetF) < 0.1 &
!sex.discrep &
!rownames(info.indv)%in%ids.rel
mask.indiv <- use & !is.na(use)
genotype.qc <- genotype.qc.snps[mask.indiv, ]
phenotype.qc <- colorectal.phenotype[mask.indiv, ]
identical(rownames(phenotype.qc), rownames(genotype.qc))
dim(phenotype)
dim
```

As before, we reported the individuals removed using the script:

```
# Number of individuals removed to bad call rate
sum(info.indv$Call.rate < 0.95)
# Number of individuals removed for heterozygosity problems
sum(abs(info.indv$hetF) > 0.1)
# Number of individuals removed for sex discrepancies
sum(sex.discrep)
# Number of individuals removed to be related with others
length(ids.rel)
# The total number of individuals that do not pass QC
sum(!mask.indiv)
```

The number of deleted individuals is 69.

4 Results

We are performing a Genome Wide Association Analysis, a GWAS, where we can see how every SNP is regressed individually on our trait of interest, in this case, the colorectal cancer. To visualise this we will need to create a Manhattan plot.

Before creating the plot, we need to create the GWAS summary statistic file, which should have the following columns: the chromosome, the position of the SNP on the chromosome, the p-value and the SNP name.

Once we have this file, we can proceed to create the Manhattan plot. This type of plot shows a point for every SNP or locus tested. The axis show the position in the genome of each SNP, on the x-axis, and the -log10 p-value, on the y-axis.

To evaluate the significance of each SNP we use the Bonferroni-corrected treshold, which is 0,05 divided by the number of SNPs in the summary statistics. Once we calculate this, in the plot we add a line using this significance, so if a SNP is placed over this line we know it has significance.

The final Manhattan plot of the GWAS is (Figure 2).

5 Discussion

Looking at the obtained Manhattan plot (Figure 2) we can see 11 SNPs which, after our analysis, are found associated to colorectal cancer. The SNPs and their chromosome location are in Table 1. These SNPs should not be immediately considered for molecular studies because before of that, the available information of them in different data bases and articles should be searched. If some SNPs are possible candidates to have relation with our cancer, they should be considered for molecular studies.

Below, each SNP is described with the information available and discussed its possible implication into colorectal cancer.

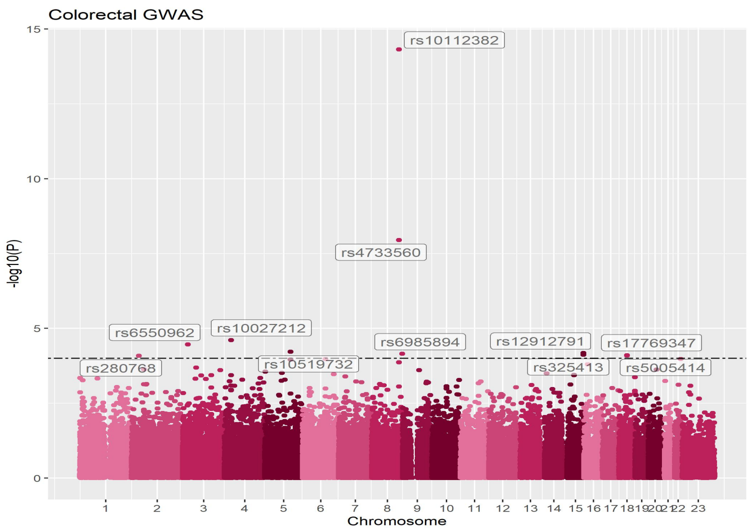

Figure 2: Manhattan plot

SNP	CHR
rs280768	2
rs6550962	3
rs10027212	4
rs10519732	5
rs4733560	8
rs10112382	8
rs6985894	8
rs12912791	15
rs325413	15
rs5005414	18
rs17769347	18

Table 1: SNPs differential expressed and their chromosome location

The SNP **rs280768** is in Chromosome 2 and any gene is reported in this region (Figure 3). C and T are the alleles of this SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.529 and 0.471, respectively.

The SNP **rs6550962** is in Chromosome 3, specifically in *RARB* (retinoic acid receptor beta) gene (Figure 4). A and G are the alleles of this SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.868 and 0.132, respectively. This SNP is reported at SO Term (Sequence Ontology Term) as Genic Upstream Transcription Variant and Intron Variant.

RARB gene encodes a nuclear transcriptional regulator and the protein is located in cytoplasm and subnuclear compartments. It binds retinoic acid and mediates cellular signalling in embryonic morphogenesis, cell growth and differentiation. Wang et al., in 2019, concluded that *RARB* expression was strongly correlated with several clinicopathological factors of colorectal cancer and may represent a favourable prognostic marker in patients with this cancer. Furthermore, Catalogue Of Somatic Mutations In Cancer (COSMIC) reports 67 samples with mutation in this gene from a total of 2340 samples. According to these results, the consequence of this SNP should be studied because it can influence in the overexpression of the gene.

The SNP **rs10027212** is in Chromosome 4, specifically in *PCDH7* (protocadherin 7) gene (Figure 5). T and G are the alleles of this SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.544 and 0.456, respectively. This SNP is reported at SO Term as Genic Downstream Transcript Variant and Intron Variant.

PCDH7 gene encodes an integral membrane protein that is thought to function in cell-cell recognition and adhesion. Wu et al., in 2018, found molecular evidences which support the vital role of a novel AQP8-PCDH7 signalling axis in growth and metastasis of colorectal carcinoma. It was already known that *PCDH7* is overexpressed in several malignancies and are correlated with cancer cells metastasis. So, they confirmed that in colorectal cancer is overexpressed too. Furthermore, COSMIC reports 125 samples with mutation in this gene from a total of 2314 samples. Hence, this SNP should be considered to be studied due to it can influence in the progression of the tumor. The SNP **rs10519732** is in Chromosome 5, specially in *CSNK1G3*

(casein kinase 1 gamma 3) gene (Figure 6). T and C are the alleles of this SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.893 and 0.107, respectively. This SNP is reported at SO Term as Intron Variant.

CSNK1G3 gene encodes a member of a family of serine/threonine protein kinases. Currently, we do not find any publication showing the relation between mutations of this gene and colorectal cancer. Therefore, this SNP should be studied after those which show possible association supported by bibliography.

Both SNPs **rs4733560** and **rs10112382** are in Chromosome 8, specially in an intergenic region (Figure 7 and 8), but in the surrounded area there are different genes such as *CASC11*, *MYC* and *PVT1.* G and A are the alleles of the former SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.740 and 0.260, respectively. On the other hand, T and C are the alleles of the later SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.339 and 0.661, respectively.

CASC11 (cancer susceptibility candidate 11) is long non-coding RNA (lncRNAs). Zang et al, in 2016, reported that this gene was upregulated in colorectal cancer (CRC) tissues and increased *CASC11* expression in CRC was associated with tumor size, serosalinvasioin, lymph metastasis and tumor-node-metastasis (TNM) stage.

MYC is a proto-oncogene encodes a nuclear phosphoprotein that plays a role in cell cycle progression, apoptosis and cellular transformation. Satoh et al., in 2017, reported that metabolic reprogramming of colorectal cancer is caused by aberrant *MYC* expression, as well as, knockdown of MYC in colorectal cancer cells reset the altered metabolism and supressed cell growth.

PVT1 is a long non-coding RNA locus that has been identified as a candidate oncogene. Increased copy number and overexpression of this gene are associated with many types of cancers. He et al., in 2019, reported that the CRC-associated lncRNA *PVT1* is a key regulator of CRC development and progression.

Taking all together, both SNPs can affect the expression of these gens due to they can be at the promotor region of them. So, this SNPs should be studied in a molecular study.

SNP **rs6985894** is in chromosome 8, as well (Figure 9). At the surrounding region of this SNP there are different genes. Some of them are *BAI1* (Brain-specific angiogenesis inhibitor 1), *ARC* (activity regulated cytoskeleton associated protein), *PSCA* (prostate stem cell antigen) and *JRK* (Jrk helix-turn-helix protein). A and G are the alleles of the SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.218 and 0.782, respectively.

BAI 1 is postulated to be a member of the secretin receptor family, an inhibitor of angiogenesis. Angiogenesis has been shown to be essential for growth and metastasis of solid tumors. Fukushima et al., in 1998, reported that the expression of this gene was significantly reduced in colorectal cancers as compared to the extraneoplastic tissues. So, *BAI1* expression was inversely correlated with vascular invasion and metastasis.

PSCA gene encodes a glycosylphosphatidylinositol-anchored cell membrane glycoprotein. This gene is up-regulated in different types of cancers. Even though this gene has been implicated in the pathogenesis of several solid tumours, Smith et al., in 2012, do not found associations between a determinate polymorphism and the risk of colorectal adenomata or cancer. They concluded that it seemed unlikely that PSCA has a role in the initiation or progression of colorectal neoplasia.

JRK gene encodes a conserved protein that is similar to DNA-binding proteins. Inactivation of the related gene in mice resulted in epileptic seizures. Currently, there are not any publication showing the relation between mutations of this gene and cancer. *ARC* gene does not have any reported relation with cancer, neither.

According to the information above, we think that this SNP has to be molecular study because it can variate the promotor activity of one of these genes and it conclude in increase risk of cancer.

The SNPs **rs12912791** and **rs325413** are in Chromosome 15, both are in *MEF2A* (myocyte enhancer factor 2A) gene (Figure 10 and 11). T and C are the alleles of the former SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.732 and 0.268, respectively. G and A are the alleles of the later SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.379 and 0.621, respectively. Both SNPs are reported at SO Term as Intron Variant, as well as Genic upstream Transcript Variant only former and Genic Downstream Transcript Variant later.

MEF2A gene encodes a DNA-binding transcription factor that activates muscle-specific, growth factor-induced and stress-induced genes. Defects in this gene could be a cause of autosomal dominant coronary artery disease 1 with myocardial infarction (ADCAD1). Despite the fact that any paper which shows a correlation between this SNP and our type of cancer is found, COSMIC reports 54 samples with mutation in this gene from a total of 2303 samples.

The SNPs **rs5005414** and **rs17769347** are in Chromosome 18 in a very close region (Figure 12 and 13). G and A are the alleles of the former SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.910 and 0.090, respectively. A and G are the alleles of the later SNV with a frequency, according to The 1000 Genomes Project with global population sample, of 0.924 and 0.076, respectively

All things considered, rs6550962 (*RARB* gene), rs10027212 (*PCDH7* gene), rs12912791 (*MEF2A* gene) and rs325413 (*MEF2A* gene) are the SNPs which should be considered for molecular studies, due to there are evidences of an association with colorectal cancer. rs10519732 (*CSNK163* gene) should be studied after because it is located in a codificant region. As well as rs4733560, rs10112382 and rs6985894 because this variation can modify the promotor activity of oncogens, such as *MYC*. Finally, the remaining SNPs, should be study the last one because they are located in a region where there is no genes annotated.

Besides this research, we have also looked for signatures characteristic of this type of cancer in order to determinate if the mutations we found through this GWAS analysis can also correlate with the most frequent signatures in colorectal cancer. This information is in a different document in ”Calaix de sastre” folder.

6 Bibliography

Fukushima, Y., et al. (1998). Brain-specific angiogenesis inhibitor 1 expression is inversely correlated with vascularity and distant metastasis of colorectal cancer. International Journal of Oncology, 13(5), 967-970. https://doi.org/10.3892/ijo.13.5.967

He, F., et al. (2019). Long noncoding RNA PVT1-214 promotes proliferation and invasion of colorectal cancer by stabilizing Lin28 and interacting with miR-128. Oncogene, 38(2), 164-179. https://doi.org/10.1038/s41388-018-0432-8

Satoh, K., et al.(2017). Global metabolic reprogramming of colorectal cancer occurs at adenoma stage and is induced by MYC. Proceedings of the National Academy of Sciences of the United States of America, 114(37), E7697-E7706. https://doi.org/10.1073/pnas.1710366114

Smith, C., et al. (2012). Lack of association between the rs2294008 polymorphism in the prostate stem cell antigen gene and colorectal neoplasia: A case-control and immunohistochemical study. BMC Research Notes, 5. https://doi.org/10.1186/1756-0500-5-371

Wang, W., et al.(2019). High expression of RAR is a favorable factor in colorectal cancer. Disease Markers, 2019. https://doi.org/10.1155/2019/7138754.

Wu, D. Q., et al. (2018). AQP8 inhibits colorectal cancer growth and metastasis by down-regulating PI3K/AKT signaling and PCDH7 expression. American journal of cancer research, 8(2), 266-279. http://www.ncbi.nlm.nih.gov/pubmed/29511597

Zhang, Z., et al. (2016). Long non-coding RNA CASC11 interacts with hnRNP-K and activates the WNT/-catenin pathway to promote growth and metastasis in colorectal cancer. Cancer Letters, 376(1), 62-73. https://doi.org/10.1016/j.canlet.2016.03.022

https://www.overleaf.com/learn/latex/Tables#Creating_a_simple_table_in_LaTeX

https://www.ncbi.nlm.nih.gov/snp/

https://www.ncbi.nlm.nih.gov/gene/

https://www.ncbi.nlm.nih.gov/pubmed/

https://cancer.sanger.ac.uk/cosmic

7 Appendix with supplementary figures

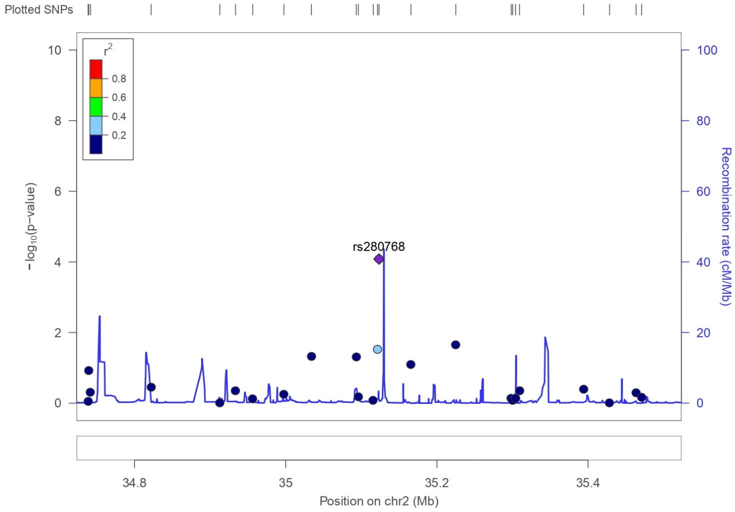

Figure 3: SNP rs280768 localization in the genome. Obtained with LocusZoom

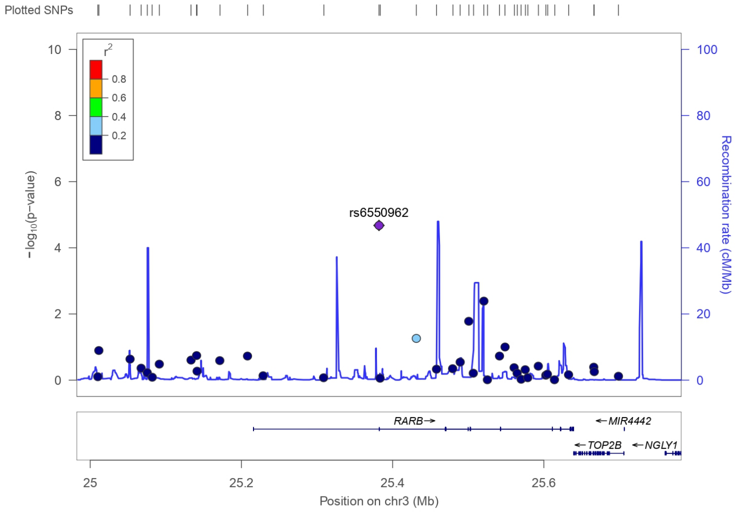

Figure 4: SNP rs6550962 localization in the genome. Obtained with LocusZoom

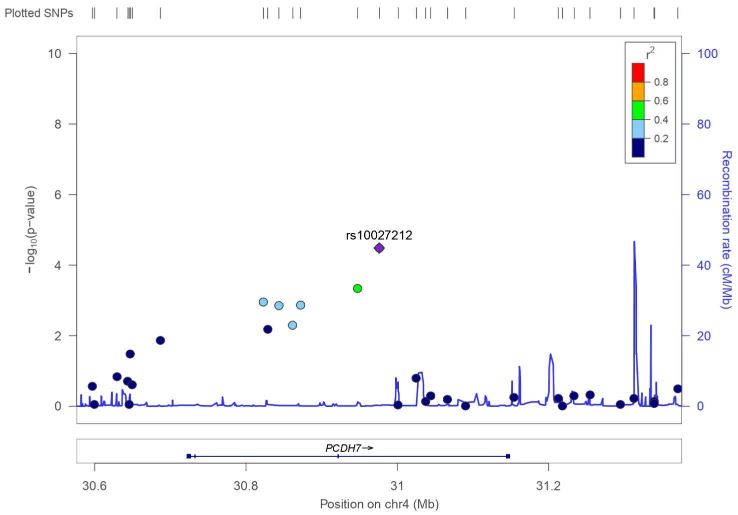

Figure 5: SNP rs10027212 localization in the genome. Obtained with LocusZoom

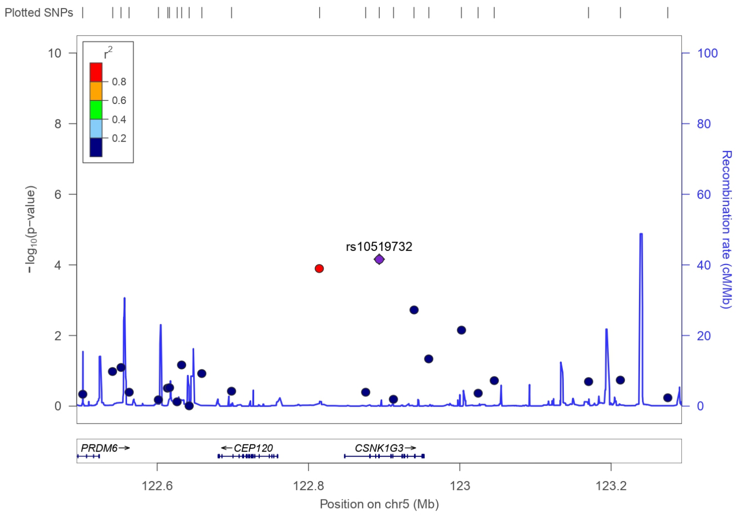

Figure 6: SNP rs10519732 localization in the genome. Obtained with LocusZoom

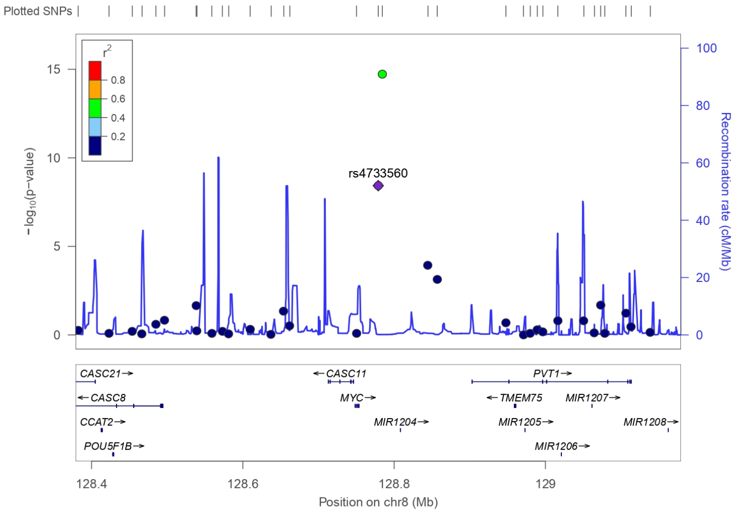

Figure 7: SNP rs4733560 localization in the genome. Obtained with LocusZoom

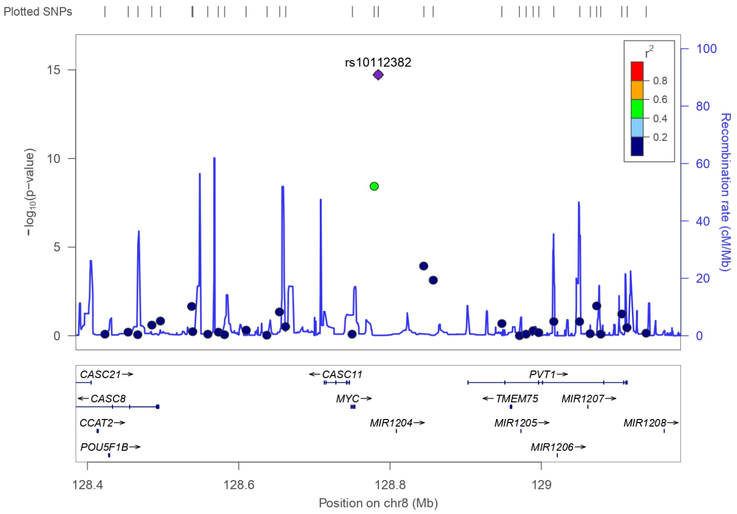

Figure 8: SNP rs10112382 localization in the genome. Obtained with LocusZoom

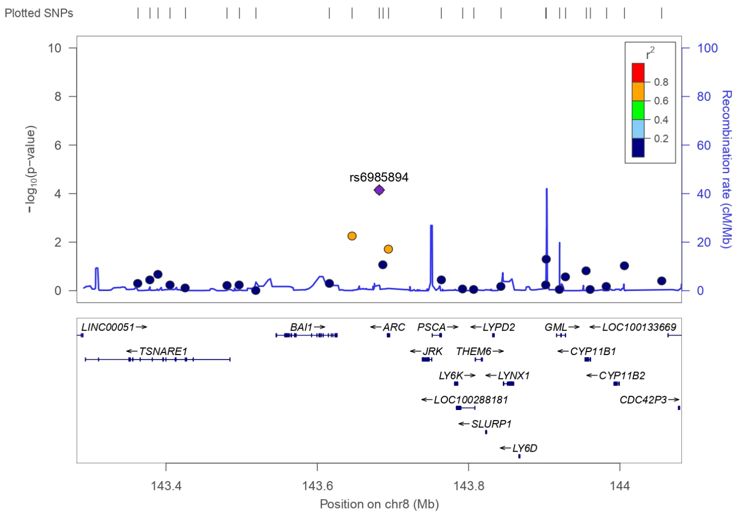

Figure 9: SNP rs6985894 localization in the genome. Obtained with LocusZoom

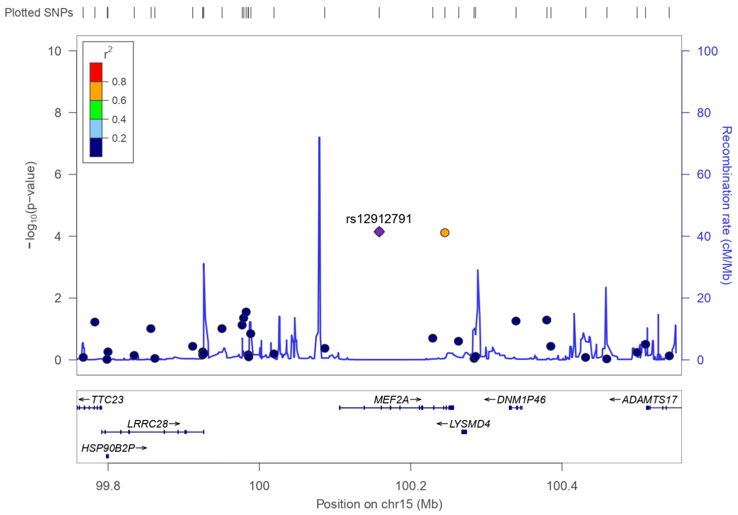

Figure 10: SNP rs12912791 localization in the genome. Obtained with LocusZoom

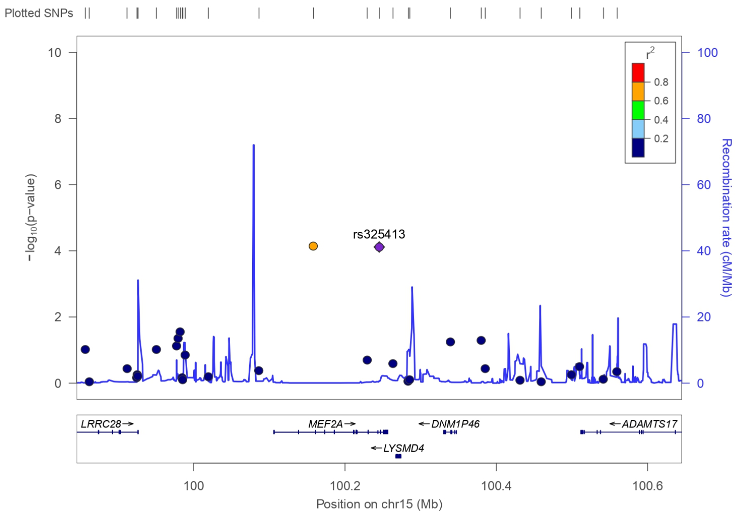

Figure 11: SNP rs325413 localization in the genome. Obtained with LocusZoom

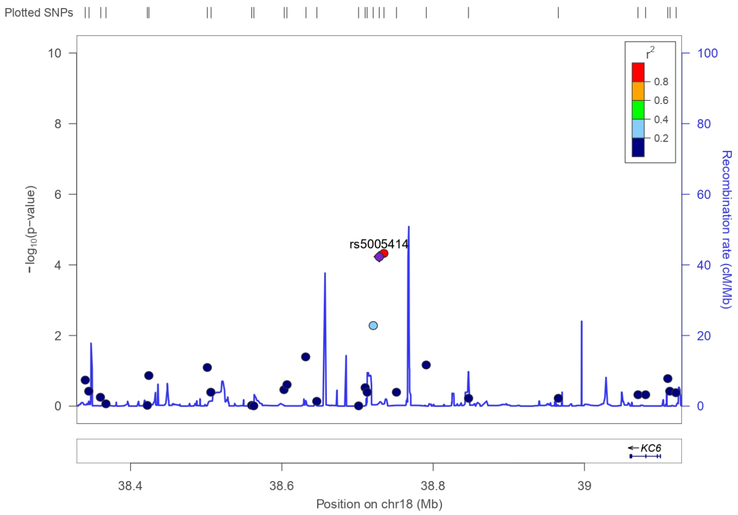

Figure 12: SNP rs5005414 localization in the genome. Obtained with LocusZoom

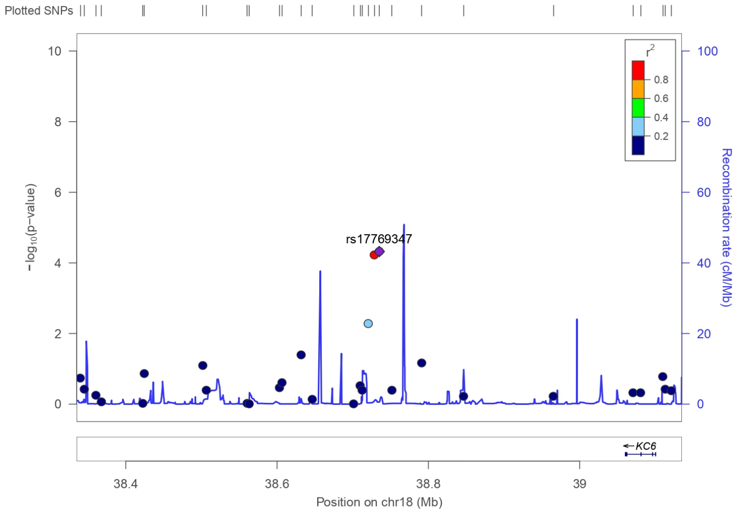

Figure 13: SNP rs17769347 localization in the genome. Obtained with LocusZoom