January 2022
Grade in Genetics 2021-22

Genome-wide association study of Colorectal Cancer

Universitat Autònoma de Barcelona

Mar Alvarez, Claudia Beneyto, Rocío Calvo and Carla Hijazo

Professors: Marta Coronado and Nerea Carron

Current Topics in Bioinformatics

Abstract: *Colorectal cancer (CRC) is a leading cause of cancer-related death worldwide and has a strong heritable basis. We report a case-control genome-wide association analysis (GWAS of 2312 subjects and has 100000 SNPs analysed. The packages used in this association study implemented for data preparation, quality control, the GWAS study itself and its representation in a Manhattan plot were ggplot2, dplyr, ggrepel, SNPassoc and BiocManager. We represented the distribution of age, gender and frequency of smoking habits to figurate better the role of this items in the colorectal cancer. Also, a quality control for SNPs and individuals was made before association testing to prevent errors. 9 novel risk loci for CRC were identified with a significant level ($P<1 \times 10^{-4}$) at the GWAS. Some variants such as s1890668 or rs1550051 were found to be in non-coding sequences, while other variants such as rs2290753 and rs11674328, were located inside the protein-coding genes such as AKT3 or HECW2.*

Keywords: **Colorectal Cancer, GWAS, SNPs**

Introduction

Colorectal cancer (CRC) is the third most common diagnosis and second deadliest malignancy for both sexes combined. CRC has both strong environmental associations and genetic risk factors. The change of the normal colonic epithelium to a precancerous lesion and ultimately an invasive carcinoma requires an accumulation of genetic mutations either somatic (acquired) and/or germline (inherited) in an approximately 10 to 15-year period [1].

Previous Family-based linkage studies and recent whole-exome sequencing studies have identified multiple CRC susceptibility genes, such as *APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *BMPR1A*, *NTHL1* and *TP533-5*. Deleterious mutations in these genes, however, are rare and account for less than 6% of CRC cases in the general population cancer [2].

Genome-wide association studies (GWAS) provide a powerful new approach to identify common, low penetrance susceptibility loci. The risk prediction ability of susceptibility markers identified in GWAS for CRC may improve as more variants are discovered. This may in turn have important implications for targeting high risk individuals for colonoscopy screening [3]. Due to the high incidence of CRC, it is of particular interest to identify new polymorphic variants that may be useful in improving prognosis and early diagnosis. In the following project we have implemented a GWAS analysis in colorectal cancer. The analysis has been run with the R software, below we will see which variants are behind colorectal cancer.

Methods

To perform this GWAS, *Rstudio* was used to facilitate data processing and analysis.

Packages and tools used

The packages used in this association study implemented for data preparation, quality control, the GWAS study itself and its representation in a Manhattan plot were:

- *ggplot2*: used to declare the input data frame for a graphic and to specify the set of plot aesthetics intended to be common throughout all subsequent layers.
- *dplyr*: is a grammar of data manipulation, providing a consistent set of verbs that help you solve the most common data manipulation challenges.
- *ggrepel*: ggrepel provides geoms for ggplot2 to repel overlapping text labels.
- *SNPassoc*: it contains facilities for data manipulation, tools for exploratory data analysis, convenient graphical facilities, and tools for assessing genetic association for both quantitative and categorical (case-control) traits in whole genome approaches. Genome-based studies are normally analysed using a multi-stage approach.
- BiocManager: it allows us to install and manage packages from the Bioconductor project. Bioconductor focuses on the statistical analysis and comprehension of high-throughput genomic data. We used a range of functions of the *snpStats* package to adjust analyses for clinical, environmental, and/or demographic factors as well as ancestral differences between the subjects. The R package *SNPRelate* is used to perform identity-by-descent (IBD) analysis, computing kinship within the sample.

Data description

In the present GWAS case-control study, we analysed 100,000 SNPs using data from 2312 subjects [5].

The data of the participants from which we proceeded was:

1. genomic SNP data (stored in Binary BED file);
2. SNP annotations such as chromosome, SNP name, position in morgans, base-pair coordinates, reference nucleotide and alternative nucleotide (stored in Text BIM file);
3. individual's family information containing family identifier, individual ID, paternal ID, maternal ID, sex and phenotypes (stored in Text FAM file) and lastly a text file used to add phenotypic information, clinical or epidemiological data.

With these data we can extract different graphs using *ggplot2*, it can be significant to know how the subjects of our study are distributed according to different variables.

Considering the total number of subjects, we represented the distribution of age and gender. As can be observed from Figure 1 that our cohort of study is gender balanced and age ranges from 30 to 55 years old, with a maximum of subjects around 40 years old. So, it could be said that our population sample data is representative for colorectal cancer patients. Secondly, because it is a considerable risk factor for cancer, we also wanted to represent the frequency of smoking habits of the population sample, show in Figure 2. Most of the subjects are current smokers but also a great number of individuals don't smoke and finally a tiny proportion of them are ex-smokers.

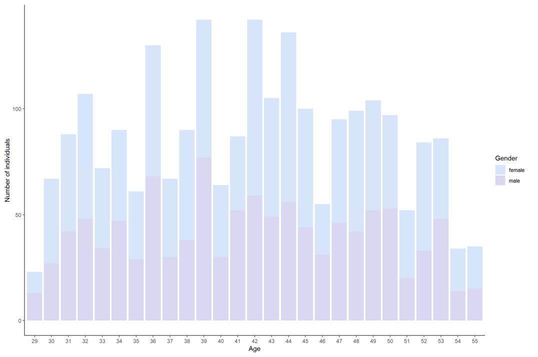

Figure 1: *Age and sex distribution. Graphic made with Rstudio.*

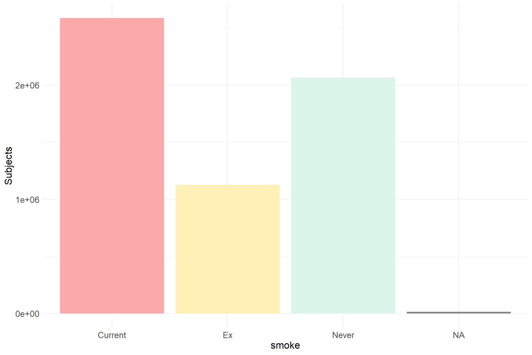

Figure 2: *Smoking habits distribution. Graphic made with Rstudio.*

Quality control

We performed the quality control (QC) of genomic data at the SNP and individual levels, before association testing.

Quality control of SNPs

To carry out SNPs QC we first removed markers with a call rate less than 95%, as SNPs with high level of missingness could lead to bias. Rare SNPs (MAF<5%) were also screened. Lastly, we filtered the SNPs of the control cohort (1138 control subjects) which did not pass de Hardy-Weinberg Equilibrium (HWE) test. The HWE was only tested in controls as the non-compliance of the HWE law in cases can be indicative of true genetic association with disease risk. For the HWE test a parsimonious threshold of 0.001 was set, corresponding to a z-score of ± 3.3.

After performing SNPs quality control, the number of SNPs removed due to a bad call rate, low MAF or violation of HWE were 875, 10669 and 72 respectively, bringing de number of SNPs removed from our association study to 11479. Therefore, out of 100000 SNPs, 88521 SNPs were left after QC of the SNPs. As mentioned above, HWE test was exclusively conducted in controls, which resulted to be 1138 subjects.

Quality control of Individuals

Regarding individuals' QC four filtration steps were followed:

1. Identification of individuals with discordant reported and genomic sex. Gender was inferred from the heterozygosity of chromosome X. Males have an expected heterozygosity of 0 and females of 0.3. The Figure 3 shows that there were some reported males with nonzero X- heterozygosity and females with zero X- heterozygosity.

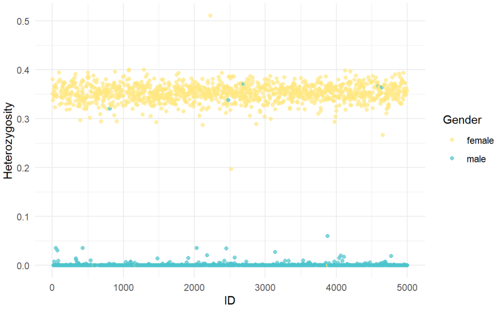

Figure 3: *Gender heterozygosity. Graphic made with Rstudio.*

This individuals were later removed from the study. The difference between the assigned sex and the sex determined based on the genotype usually points to sample mix-ups in the lab.

2. Identification of individuals with outlying missing genotype or heterozygosity rate, meaning the proportion of heterozygous genotypes. High levels of heterozygosity within an individual might be an indication of low sample quality.

Heterozygosity was computed from the statistic:

$$F = 1 - \frac{f(Aa)}{E(f(Aa))}$$

where the numerator is the observed proportion of heterozygous genotypes (Aa) of a given individual and the denominator is the expected proportion of heterozygous genotypes, the last one was computed from the MAF (Minor Allele Frequency) across all the subject's non-missing SNPs. See Figure 4

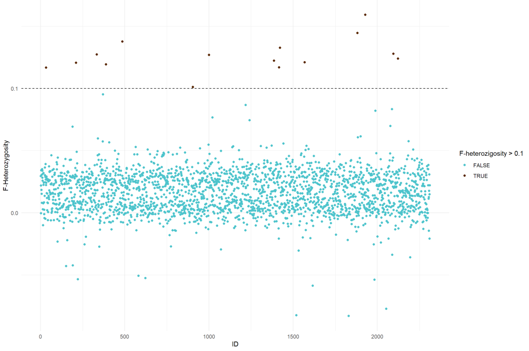

Figure 4: *Heterozygosity F-statistics. Graphic made with Rstudio.*

3. Identification of duplicated or related individuals and, identification of individuals of divergent ancestry from the sample. As GWAS assumes that all subjects are genetically unrelated, we performed the appropriate correction in order not to include relatives in the association study, which could lead to biased estimations of standard errors of SNP effect sizes. We searched individuals whose relatedness was higher than expected and removed adjacent SNPs that exceed a Linkage Disequilibrium threshold. A pair of individuals with higher-than-expected relatedness are considered with kinship score > 0.1. Those related individuals were later removed from the study.

Summing up, individuals with more than 5% missing genotypes, with sex discrepancies, F-heterozygosity absolute value > 0.1 and kinship coefficient > 0.1 were removed from the genotype and phenotype data.

After performing individuals' quality control, we kept 2252 individuals from 2312. The number of individuals removed due to a bad call rate, heterozygosity problems and relatedness were 32, 15 and 15 respectively. There were no sex discrepancies.

Once the quality control had been completed, we proceeded to carry out the Genome Wide Association Study and create the Manhattan plot (GWAS visualization) afterwards. For the genome-wide significant line we chose to get a Bonferroni-corrected threshold (P<0.0001).

Results

As shown in Figure 5, we identified 9 novel risk loci for CRC at the genome-wide significance level (P<0.0001); rs2290753, rs11674328, rs1890668, rs786319, rs1550051, rs7905846, rs10083549, rs8080301 and rs926331. We used LocusZoom [4] to find if either SNPs overlapped protein-coding genes. See Appendix for the overlapping variants obtained with LocusZoom, Figures: 6, 7, 8, 9, and 10.

The variants rs1890668, rs1550051, rs7905846, rs926331 were found to be in non-coding sequences, while the variants rs2290753, rs11674328, rs786319, rs10083549, rs8080301, were located inside the protein-coding genes *AKT3*, *HECW2*, *PRUNE2*, *GABRG3*, *RAP1GAP2*. See Manhattan plot in Figure 5.

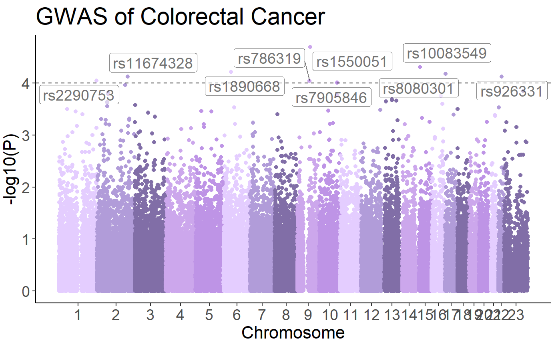

Figure 5: *Manhattan plot. Graphic made with Rstudio.*

Discussion

It is important to be able to identify individuals at high risk of CRC to enable enhanced screening and other interventions. Equally pressing is the need to identify individuals at low risk to prevent unnecessary screening and associated complications.

With the aim of founding if these variants and their overlapping genes had been reported before, we used Ensembl [6] to look for phenotypes or traits associated with the gene and variant. We also checked out pathology from Human Protein Atlas [7]. Furthermore, we sought articles that associated the variant or gene with colon, rectal or colorectal cancer in PubMed [1].

Lastly, we used Varsome [8], which can be used as a bioinformatic predictor and classifies variants as benign, potentially benign, uncertain significance, potentially malignant or malignant.

All the significant loci overlapping coding sequences obtained in the GWAS, were found to be intron variants. Follow up we will proceed to discuss each locus and its characteristics.

rs2290753

The risk associated with locus rs2290753, that it is inside AKT3, has not been reported before, and Varsome predicts it's as a benign variant. However, there is a relation between the gene and colorectal cancer, seen both in ensembl and PubMed. Ensembl shows an association with colon and colorectal adenocarcinoma and other kind of tumours.

The Akt genes act as nodal oncoproteins which seems to drive cell survival mechanisms that contribute to cancer progression and metastasis. Thus, targeting Akt as a potential anti-metastatic therapeutic strategy either as a single agent or in combination holds significant promise [9]. Especially, AKT3 expression in Mesenchymal Colorectal Cancer Cells has been associated with Epithelial-Mesenchymal Transition [10]. Another article has reported the potential prognosis and diagnostic value of AKT3 [11].

rs11674328

There is no information about this locus (overlapping HECW2) and Varsome qualifies the locus as a variant of uncertain significance. Whereas Ensembl associates the HECW2 with neurodevelopmental disorder with hypotonia, seizures, absent language with intellectual disability, neuroblastoma, cholesterol, and body fat distribution, but not with colorectal cancer. No articles have been found linking this gene to colorectal cancer. In fact, this gene is not very studied because there is practically not enough bibliography. According to Human Protein Atlas, there is a linkage between this locus and renal cancer.

rs786319

This variant is predicted to be benign although there are no articles published. PRUNE2 is associated with body heigh, fat mass, hippocampal atrophy and blood pressure, but no with colorectal cancer. However, this article states that "PRUNE2 is a human prostate cancer suppressor" [12].

rs10083549

This variant overlapping GABRG3 gene is also predicted to be a benign variant, but there is a lack of more information. GABRG3 has been associated with blood pressure, type 2 diabetes and ovarian cancer. According to Human Protein Atlas the gene is related with prostate cancer and with distinct diagnostic and prognostic values of γ-aminobutyric acid type A receptor (GABRG) family genes in patients with colon adenocarcinoma [13].

rs8080301

This variant of uncertain significance and its overlapped gene RAP1GAP2 don't show colorectal cancer association. Ensembl associate this gene with traits or phenotypes such as asthma, blood pressure, hypertension, eosinophil and lymphocyte count.

Conclusion

This GWAS has made possibly to verify the close relationship between the AKT3 and GABRG3 genes. However, the rest of the varinats that have been referenced as signicants in other studies such as APC, there have been not reported in our GWAS. Another unexpected fact is that, there hasn't been found any relation between the genes associated with colorectal cancer in this GWAS. All this facts, leads us to believe that the sample population was not big enough to make a good prediction. The lack of patients and controls has not allowed us to correctly distinguish between significant SNPs. In conclusion, this topic requires further studies and huge GWAS.

References

[1] Recio-Boiles A, Cagir B. Colon Cancer. In: StatPearls. Treasure Island (FL): StatPearls Publishing; January 25, 2021.
`https://pubmed.ncbi.nlm.nih.gov/29262132/`

[2] Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer.
Lu Y, Kweon SS, Tanikawa C, et al.
Gastroenterology. 2019;156(5):1455-1466.
doi:10.1053/j.gastro.2018.11.066
`https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6441622`

[3] Genome-wide association studies and colorectal cancer
Le Marchand L.
Surg Oncol Clin N Am. 2009;18(4):663-668.
doi:10.1016/j.soc.2009.07.004
`https://pubmed.ncbi.nlm.nih.gov/19793573/`.

[4] LocusZoom - Plot with Your Data
`http://locuszoom.org/genform.php?type=yourdata`
Accessed 2022-01-23

[5] Large-scale Genome-wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer
Yingchang Lu, Sun Seog Kweon et al.
Gastroenterology, 156, 5, 4 2019

[6] Search - Homo_sapiens - Ensembl genome browser 105
`https://www.ensembl.org/Multi/Search/Results`
Accessed 2022-01-23

[7] The Human Protein Atlas `https://www.proteinatlas.org/`
Accessed 2022-01-23

[8] VarSome The Human Genomics Community
`https://varsome.com/`
Accessed 2022-01-23

[9] Cell Survival and Metastasis Regulation by Akt Signalling in Colorectal Cancer
Ekta Agarwal, Michael G. Brattain et al.
Cellular signalling, 25, 8, 8 2013

[10] AKT3 Expression in Mesenchymal Colorectal Cancer Cells Drives Growth and Is Associated with Epithelial-Mesenchymal Transition
Buikhuisen, Joyce Y. Barila, Patricia M.Gomez et al.
Cancers, 13, 4, 2 2021

[11] Potential Prognosis and Diagnostic Value of AKT3, LSM12, MEF2C, and RAB30 in Exosomes in Colorectal Cancer on Spark Framework
Jue Wang, Sheng Wu et al.
Journal of healthcare engineering, 2021, 12 2021

[12] PRUNE2 is a human prostate cancer suppressor regulated by the intronic long noncoding RNA PCA3
Ahmad Salameh, Alessandro k. Lee et al.
Proceedings of the National Academy of Sciences of the United States of America, 112, 27, 7 2015

[13] Distinct diagnostic and prognostic values of γ-aminobutyric acid type A receptor family genes in patients with colon adenocarcinoma
Ling Yan, Yi Zhen Gong et al.
Oncology letters, 20, 1, 7 2020

Appendix with supplementary

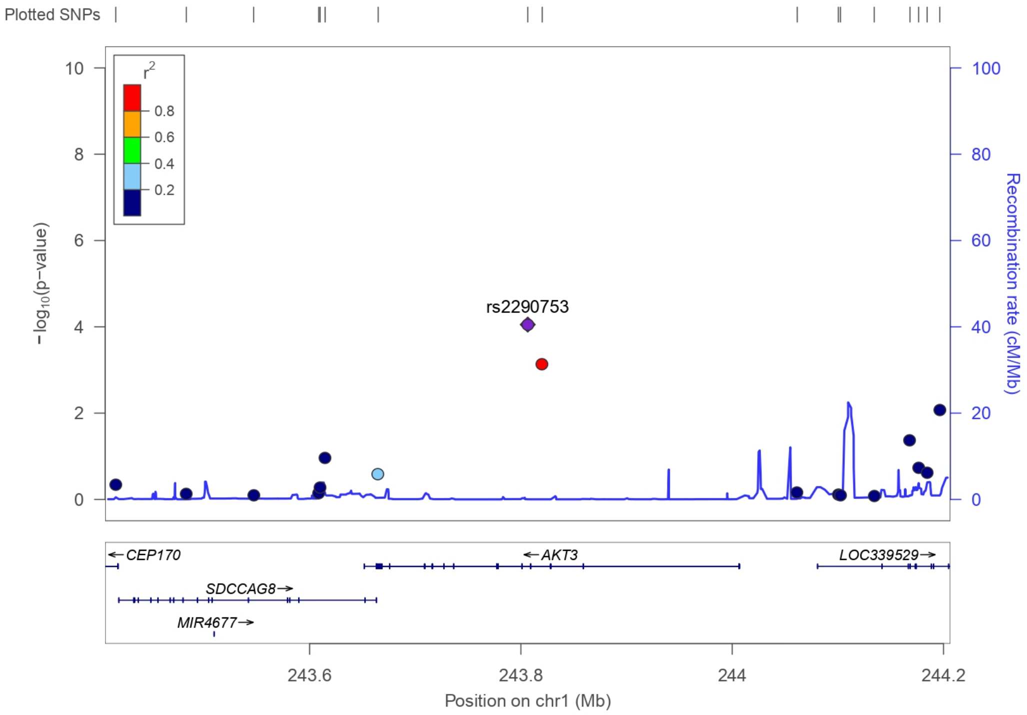

Figure 6: *rs2290753 overlapping with gen AKT3. Graphic obtained with LocusZoom.*

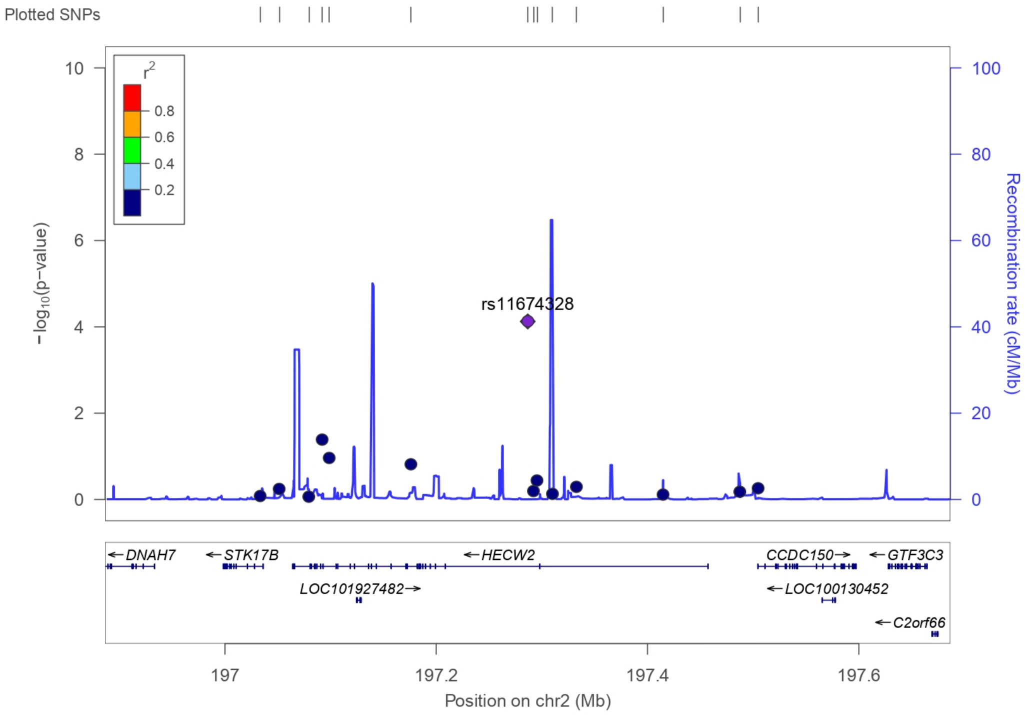

Figure 7: *rs11674328 overlapping with gen HECW2. Graphic obtained with LocusZoom.*

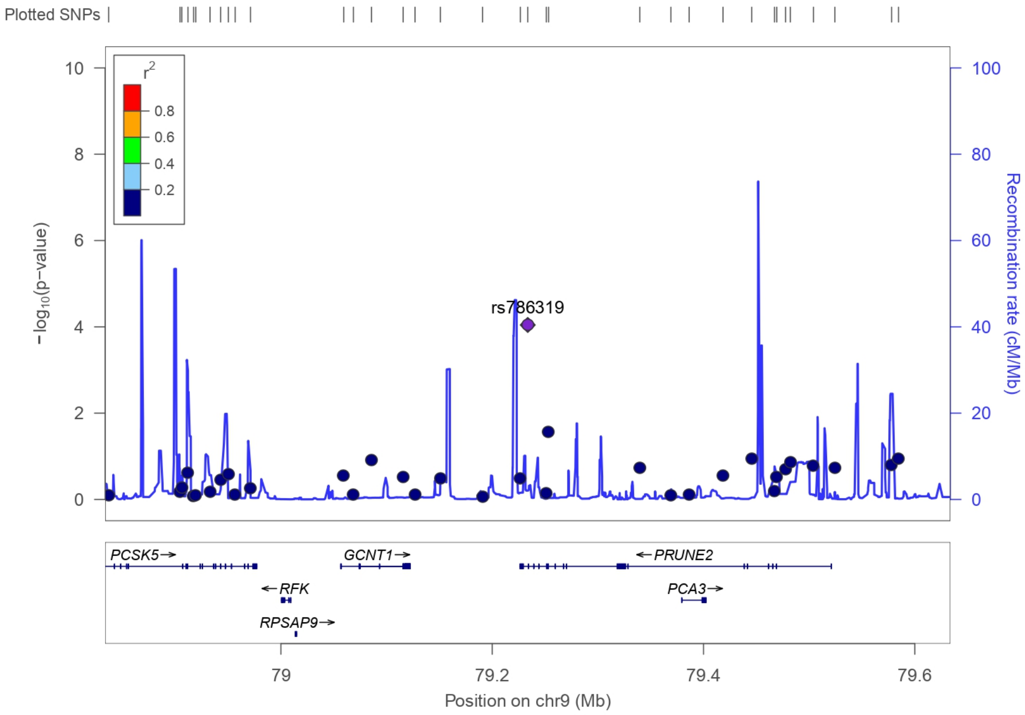

Figure 8: *rs786319 overlapping with gen PRUNE2. Graphic obtained with LocusZoom.*

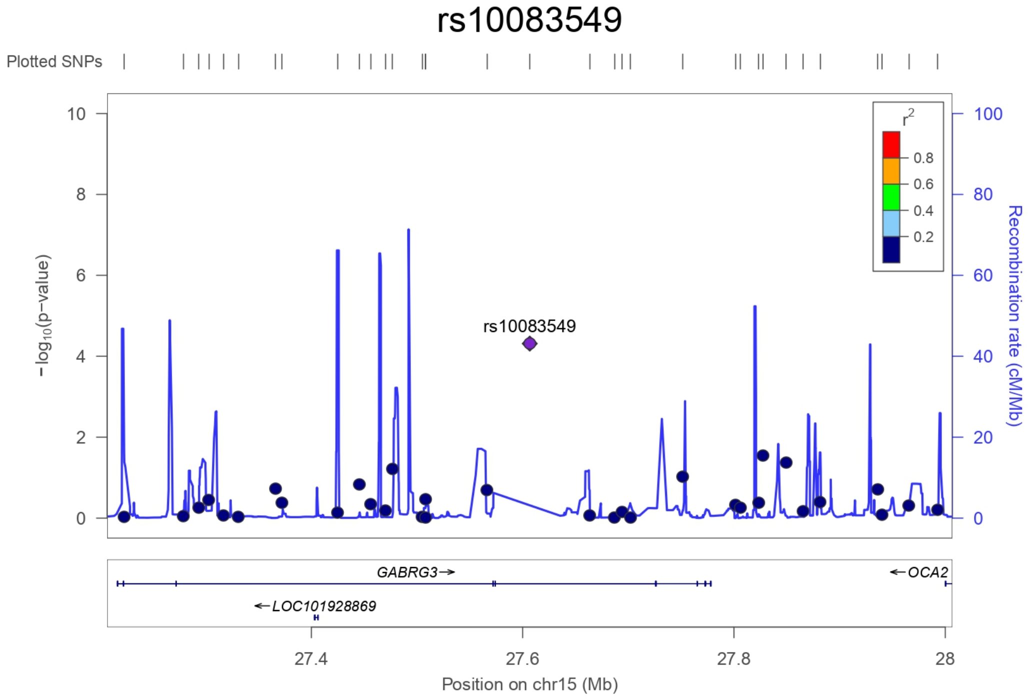

Figure 9: *rs10083549 overlapping with gen GABRG3. Graphic obtained with LocusZoom.*

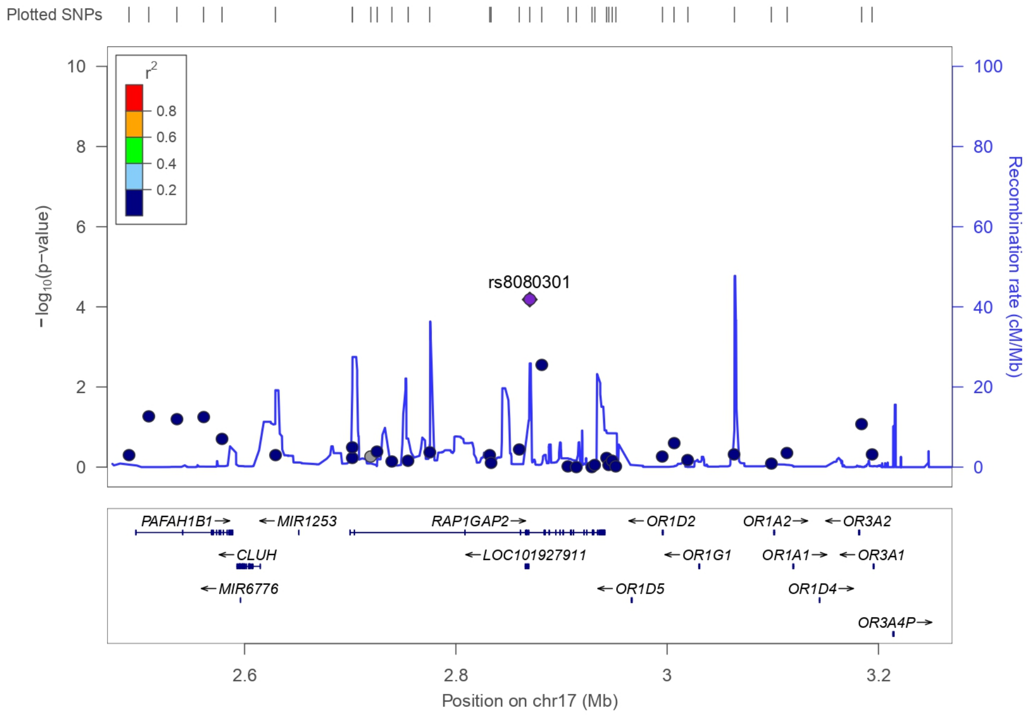

Figure 10: *rs8080301 overlapping with gen RAP1GAP2. Graphic obtained with LocusZoom.*