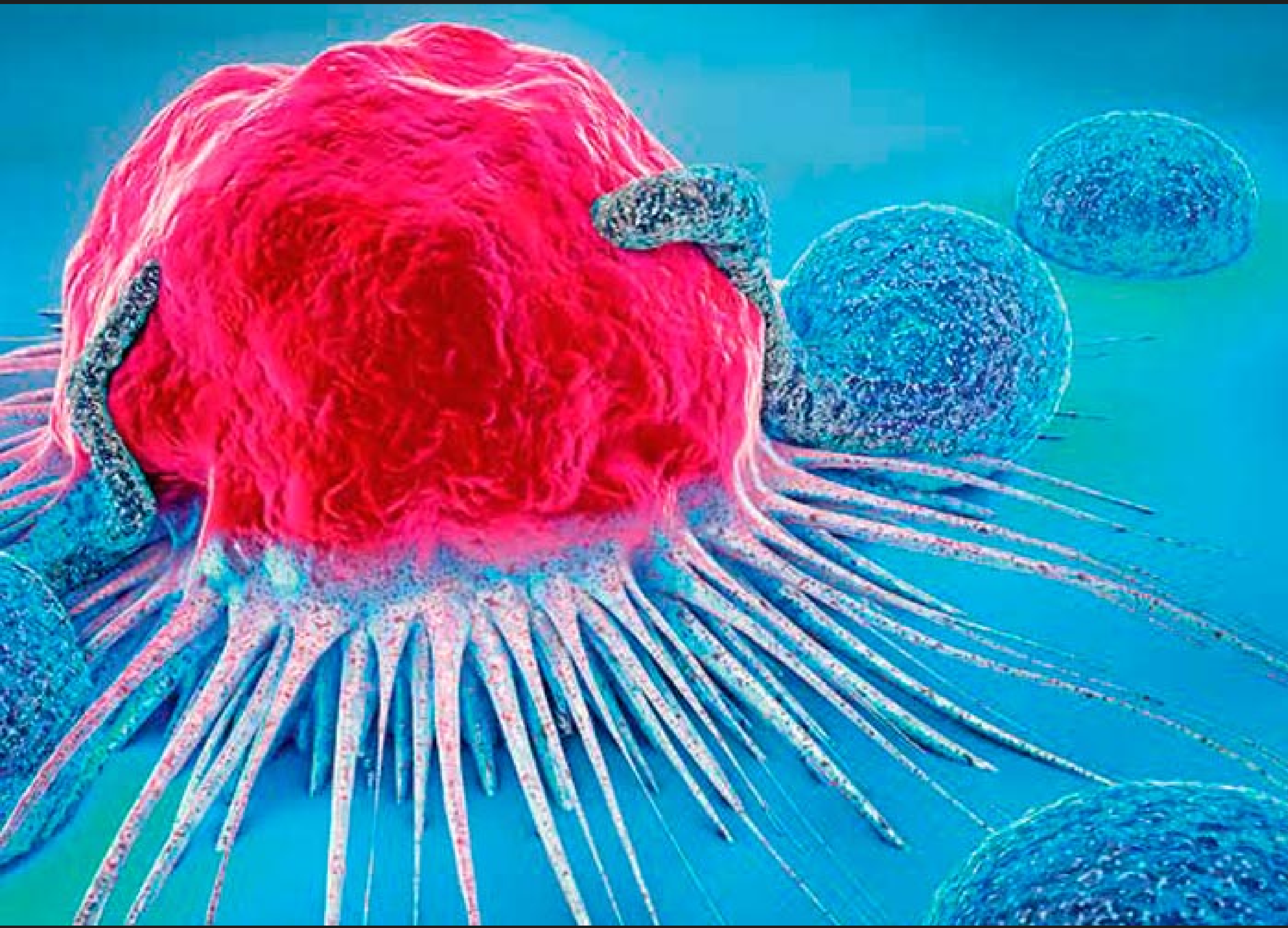
european cancer ORGANISATION

CGJ

Cancer Genetics Journal

January 2022 · Volume 36 · Issue 2



EJC EUROPEAN JOURNAL OF CANCER

cgj.bmj.com

BMJ British Medical Journal

Genome-wide Association Analysis of Colorectal Cancer Susceptibility

E. Navarro-Garre,[1,2]⋆ P. Bonet-Suñé,[1,2] A. Carrasco-Parrón,[1,2] A. Comas-Albertí[1,2] S. Redondo-Moreno[1,2]

[1] *Departament de Genètica i de Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain*
[2] *Group 5, Current Topics in Bioinformatics*

Accepted: 17 Jan 2022. Revised: 23 Nov 2021. Received: 10 Sept 2021.

Abstract
Colorectal cancer (CRC) is the fourth most common cause of cancer-related death worldwide. Its pathogenesis involves multiple environmental factors and a strong hereditary basis as approximately 25 per cent of patients have been reported to have a genetic predisposition for it. Here we perform a genome-wide association study (GWAS) to identify risk variants which can influence the pathophysiology of CRC processes. The studied population consists of 1174 CRC cases and 1138 controls which provided the identification of SNPs at 35 novel risk loci. Some of these loci are reported to have a role in other cancers so their association stands as an opportunity to pinpoint novel pathways linked to colorectal tumor generation.

Key words: Colorectal cancer – GWAS

1 INTRODUCTION

Genome-wide association studies (GWAS) allow the identification of several loci variants which occur more frequently in affected individuals with a disease than in healthy people. These polymorphisms are then inferred to have a potential role in pathogenesis and undergo posterior analyses. The development of high-quality statistical tests and whole-genome sequencing techniques grant the reliable identification of predisposition SNPs to complex conditions and diseases (Coronado and Carrón 2021).

Colorectal cancer (CRC) is responsible for one of the most frequent cancer related deaths in the world. It manifests in hosts as crypt lesions, adenomatous polyps, and carcinomes after the inactivation of adenomatous polyposis coli (APC) and other mutations. Several factors determine its incidence and geographic variation, the environment being responsible for most sporadic cases (Law et al. 2019). Nonetheless, hereditary components have an estimated contribution of 35 per cent (Le Marchand 2009) definitely play a role since 1 in 4 patients show a certain genetic predisposition (Shaik et al. 2015), In order to characterize new genetic variants contributing to CRC etiology, we perform a GWAS analysis including a yet unreported population dataset. We examine possible SNP associations underlying CRC phenotype by using R and Bioconductor packages after controlling for different variables. Finally, we underpin newly identified loci to be candidates for heritable risk of colorectal cancer.

2 METHODS

Steps and programming protocols have been strictly followed from the practice 3 script using R (Coronado and Carrón 2021). This script and other resources are available at the Data availability section of this paper.

2.1 Packages and tools

Data preparation, quality control and analysis linked to this GWAS were performed in silico using open-source Bioconductor packages (Morgan 2021) in R statistical programming language. These packages are a useful tool for large SNP association studies allowing for uncertainty in genotypes and reproducible results with powerful resolution. Moreover, devtools collection's SNPassoc package was used. Some additional R packages for data frame manipulation and data visualization were operated. These are recurrently used packages such as ggrepel (Slowikowski 2021), ggplot2 (Wickham 2016), and dplyr (Wickham et al. 2018).

- **snpStats**: classes and statistical methods for large SNP association studies. It includes generic functions to extract values from the SNP association test objects returned by various testing functions (Clayton 2021).
- **SNPRelate**: binary format for SNP data using GDS data files, a format which offers the efficient operations specifically designed for integers with two bits. It is also designed to accelerate computations on relatedness analysis using Identity-By-Descent (IBD) measures (Zheng et al. 2012).
- **SNPassoc**: perform most of the common analysis in GWAS, including descriptive statistics, missing values, calculation of Hardy-Weinberg equilibrium, analysis of association based on generalized linear models, and analysis of multiple SNPs. Permutation tests are also implemented (Gonzalez et al. 2012).

2.2 Data description

The available population dataset is public and presented in two different format files. Plink data can be loaded into R using snpStats (Clayton 2021). Genotypes are stored in a matrix where individuals are distributed in rows and SNPs in columns (alleles coded as 0, 1 or 2), and family information is contained

⋆ emiliano.navarro@autonoma.cat

in a dataframe object for every individual. SNP annotation is also stored in a dataframe object. Second file corresponds to additional phenotype information which must be merged with individual genotype data in order to get a complete characterization of this dataset.

The studied population consists of 2312 sampled individual genotypes at the start of the analysis, 1174 CRC cases and 1138 controls. Also, a number of 100,000 annotated SNPs was considered for the scope of this study. An initial description of the cohort was performed calling distribution plots for sex, age, smoking habits and body mass index (Supplementary Figure 1). The subsequent quality control (QC) analysis was conducted to filter out missing and low-quality data from both individuals and SNPs datasets.

2.3 Quality control

Previous to test associations between CRC phenotypes and genetic variants, it is crucial to perform a quality control (QC) analysis to filter out data causing bias in the final results. We conducted a set of different measures to check data quality (Coronado and Carrón 2021).

SNPs QC Genetic polymorphisms must meet a list of requirements in order to get reliable results and avoid bias caused by outlier variants. SNPs which did not meet one or more of these requirements were dismissed from analysis. Variants relevant for the study have a call rate superior to 95 per cent for the whole dataset, a minor allele frequency (MAF) superior to 5 per cent, and Hardy-Weinberg equilibrium (HWE) status in control individuals considering a significance threshold of 0.001 for being rejected. After processing the annotated data, 875 SNPs were rejected due to a bad call rate, 10,669 for low MAF, and 72 did not pass HWE testing. From the original 100,000 SNPs, 11,479 were filtered out and 88,521 remained valid for the analysis.

Individuals QC Individual samples also need filtering in order to leave out outlier subjects who can potentially skew the results or those whose biological characteristics were wrongly annotated. First, we started with the removal of individuals with discrepancies between reported and genomic sex. Genomic sex was inferred from chromosome X heterozygosity and compared with annotated gender. Expected heterozygosity for males and females is 0.0 and 0.3, respectively. Using this parameter, 9 individuals were flagged to have sex discrepancies (Figure 1).

Next we identified individuals with outlying heterozygosity from the overall genome heterozygosity rate in the dataset. Heterozygosity for every subject was computed using a F statistic.

$$F = 1 - f(Aa)/E(f(Aa))$$

where f(Aa) is the observed proportion of heterozygous genotypes of a given subject and E(f(Aa)) is the expected proportion of heterozygous genotypes, computed from the MAF across all the subject's non-missing SNPs (Coronado and Carrón 2021). After comparing F-statistic with overall heterozygosity values, individuals with a heterozygosity rate lower than 0.32 (difference bigger than 0.1) were considered outliers and therefore rejected (Figure 2).

As a means to address accurate population representation from the samples, we searched for individuals with high familial relatedness. An identity-by-descent (IBD) analysis was performed using SNPRelate (Zheng et al. 2012) package, computing kinship by its snpgdsIDMoM function for every pair of subjects. Those pairs with a kinship score bigger than 0.1 were considered to be more related than expected and thus dismissed. Last, individuals were also filtered by their call rate, rejecting those with a call rate lesser than 95 per cent.

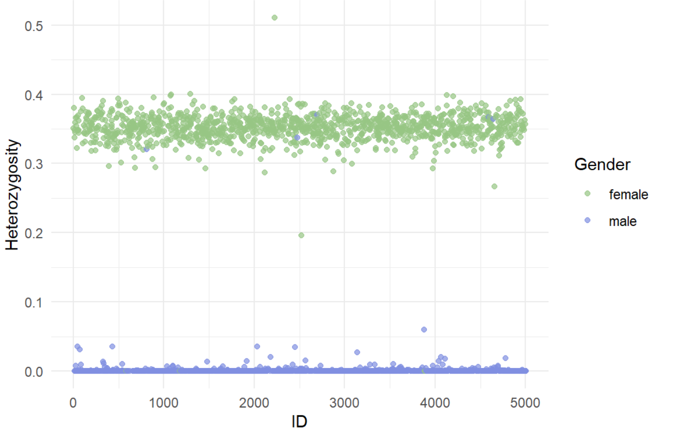


Figure 1. Chromosome X estimated heterozygosity for individuals (ID). Dot color represents individual reported gender.

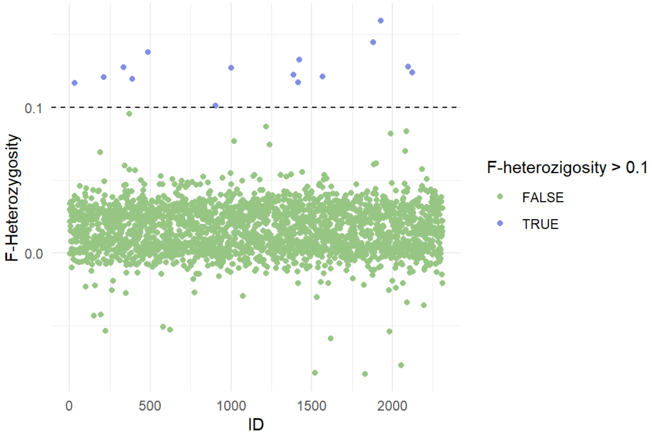


Figure 2. Comparison between F-statistic and heterozygosity for every individual in the dataset. Subjects with heterozygosity rates lower than 0.32 have a value difference bigger than 0.1.

From the 2312 individuals in the cohort, 2243 remained apt for the study while 69 were dismissed because of bad call rate (32), heterozygosity problems (15), sex discrepancies (9) and/or high relatedness (15).

3 RESULTS

3.1 Association analysis

The study protocol and regression design for this cohort can be followed through the practice 3 script (Coronado and Carrón 2021) and replicated using functions from snpStats package (Clayton 2021). After performing the described quality control procedure, we conducted genome-wide association analyses (GWAS) adjusting for different covariates (Figure 3, Supplementary Figures 2, 3, 4).

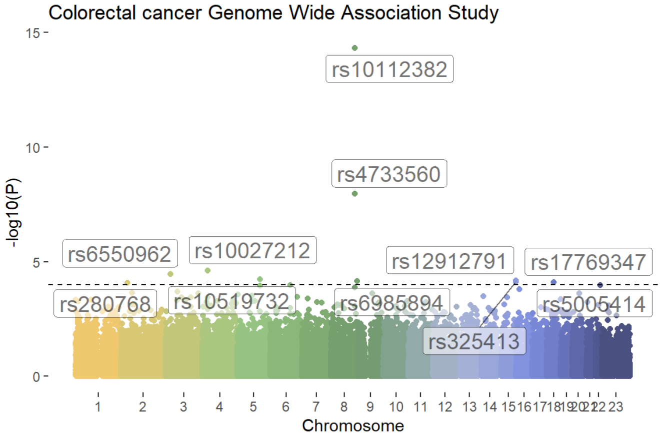


Figure 3. Manhattan plot showing the association p-values for the studied SNPs. The dashed line represents a significance threshold of P= 5 × 10-8. Loci showing a significant association are labeled.

Table 1. Summary of found SNPs with significant associations with CRC phenotype after adjusting for different covariates.

SNP	Chromosome	Position (bp)	P-value	Overlap
rs10112382	8	128853579	4E-15	-
rs4733560	8	128848183	1E-05	-
rs10027212	4	30585306	2E-05	PCDH7
rs6550962	3	25356968	3E-05	RARB
rs10519732	5	122921459	6E-05	CSNK1G3
rs12912791	15	97975967	6E-05	-
rs6985894	8	143678903	7E-05	-
rs325413	15	98062819	7E-05	MEF2A
rs17769347	18	36989057	7E-05	-
rs5005414	18	36982793	8E-05	-
rs280768	2	34976927	8E-05	-
rs9320236	6	108196381	1E-04	SCML4
rs7905846	10	115840166	8E-04	-
rs10935027	3	133863442	4E-03	NPHP3
rs1890668	6	47810527	3E-02	-
rs2346177	2	46495753	7E-02	-
rs1530621	2	46542192	0.101	-
rs6532037	4	89087979	0.146	-
rs8080301	17	2816668	0.175	RAP1GAP2
rs1888371	14	19664229	0.190	-
rs2290753	1	241873061	0.235	AKT3
rs8016597	14	19702303	0.317	-
rs2150328	14	19665092	0.413	-
rs9869834	3	63456151	0.416	SYNPR
rs786319	9	78423453	0.420	PRUNE2
rs12422767	12	128241376	0.460	TMEM132D
rs7207863	17	54892557	0.478	LINC01476
rs11674328	2	196994859	0.541	HECW2
rs6012846	20	48170921	0.556	PEDS1
rs10083549	15	25280318	0.592	GABRG3
rs1517033	18	55088717	0.711	RAX
rs7810486	7	142717233	0.760	-
rs2505115	10	30438797	0.857	JCAD
rs926331	22	35840018	0.921	-
rs1550051	9	82280085	0.990	-

We identified a total of ten loci with significant association (i. e., P < 5 × 10-8) with CRC for this cohort. An additional number of risk loci was identified when controlling for quantitative and qualitative covariates. Adjusting for age resulted in fourteen extra associated loci, and nine when controlling for BMI (Supplementary Figures 2, 3). Only two novel risk loci were added when adjusting for smoking habits: rs9320236 and rs6985894 (Supplementary Figure 4).In summary, the analysis pinpointed thirty-five SNPs (Table 1) with potential association to colorectal cancer phenotypes. However, it is important to note that only those obtained in the initial association with no trait adjustments are the most significant.

3.2 Loci region study

After identifying significantly associated markers, we searched their genomic regions in order to identify genes affected by polymorphism variations and establish a possible relationship between their functions and CRC phenotype. Several SNPs were located inside gene introns and none in either exons or regulatory regions (Table 1). Out of the original ten loci, four of them lied inside protein-coding genes (Supplementary Figure 5): PCDH7 (protocadherin), RARB (retinoic acid receptor), CSNK1G3 (casein kinase) and MEF2A (mads box transcription enhancer factor). The marker rs325413 overlapping MEF2A is also very close to LYSMD4 and DNM1P46 genes coded in the negative strand. The characteristics and relevant literature associated with these loci was consulted by using OMIM (http://www.omim.org) and dbSNP (https://www.ncbi.nlm.nih.gov/snp/) open databases.

4 DISCUSSION

Research studies conducted to this day have successfully identified 61 loci linked to CRC risk in Asian and European populations (Law et al. 2019) (Tanikawa et al. 2018). We performed a new GWAS including 1174 CRC cases and 1138 controls after a standard quality control for this dataset. Filtered SNPs and individuals which met the requirements were used for a logistic regression that estimated risk associated with CRC status.

After controlling for one qualitative trait (smoking habits) and two quantitative covariates (age and BMI), we identified 35 novel loci potentially linked to CRC susceptibility for this population (Table 1). None of them have been reported yet in previous studies nor they have been flagged as clinically relevant in any database. For the means of discussing relevant associations, we focus our analysis on the ten markers detected by the GWAS without covariate adjustments.since their significance levels are not borderline.

Based on the study of the genomic regions comprising significant loci, our data indicated several candidate genes with functions previously linked to other cancers and tumorigenesis, in particular PCDH7 (4p15.1). This gene is a protocadherin functioning in cell to cell recognition and signal transduction which harbors susceptibility for brain and breast cancer (Chen et. al. 2016). It promotes the production of cytokines and tumor necrosis factors that activate growth and chemoresistance. The expression of this gene is lower at other tissues other than heart or brain but its role in CRC should be further studied. Another locus with a relation to cancer is RARB (3p24.2), a retinoic acid receptor associated with several syndromic pathologies and also with liver cancer (De Thé et. al. 1987) when hepatic cells are infected by Hepatitis B virus. DNA integration of the virus in this site promotes over-expression of RARB and eventually tumorigenesis. Also, one of the SNPs with significant association when taking age into account (rs7810486) is suggested to have a link to marginal zone lymphoma (Lan et. al. 2009).

Some of these risk markers were mapped to regions not connected to colorectal or other cancerous processes. The gene CSNK1G3 (5q23.2) is a casein kinase with no pathological descriptions to date so it is unlikely it has any direct association with CRC phenotype. Last, MEF2A (15q26.3) is part of a myocyte-specific enhancer family that binds to muscle-specific genes with no cancer-related literature. Nonetheless, it was reported to have a booster role in smooth muscle cell proliferation (Wang et. al. 2003) hinting a possible play in cell growth. Interestingly, this locus and one of the significant SNPs when adjusting for age (rs2346177) are especially associated with heart coronary disease (Simino et. al. 2013). None of LYSMD4 and DNM1P46 have a known association with cancer.

After obtaining these GWAS results, we suggest and hope to perform future additional analyses to narrow down the candidate regions to a smaller set of markers before conducting molecular experimentation. After this, we put forward an expression quantitative trait locus (eQTL) analysis in affected samples as we sought to deepen into the real number of risk loci and the physiological mechanisms responsible for these associations.

5 FUNDING

This work was supported by no one but its authors. Fortunately, it had a cost of zero for them.

6 DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here (zipped PLINK data) and here (phenotype additional text file).
Steps of the analysis and partial results can be found in the P3 script by Coronado and Carrón.
Code for this GWAS analysis in R is available for download.

REFERENCES

Chen, Q., Boire, A., Jin, X. et. al. (2016). Carcinoma–astrocyte gap junctions promote brain metastasis by cGAMP transfer. Nature. 533, 493–498.

Clayton, D. (2021). snpStats: SnpMatrix and XSnpMatrix classes and methods. R package version 1.44.0.

Coronado, M., Carrón, N. (2021). Practice 3: Genome-Wide Association Studies (GWAS). Current Topics in Bioinformatics.

De Thé, H., Marchio, A., Tiollais, P. et al. (1987). A novel steroid thyroid hormone receptor-related gene inappropriately expressed in human hepatocellular carcinoma. Nature. 330, 667–670.

Gonzalez, J. R., Armengol, L., Solé, X., Guinó, E., Mercader, J. M., Estivill, X., Moreno, V. (2007). SNPassoc: an R package to perform whole genome association studies. Bioinformatics. 23(5), 654-655.

Lan, Q., Morton, L. M., Armstrong, B., Hartge, P., Menashe, I., Zheng, T., Purdue, M. P., Cerhan, J. R., Zhang, Y., Grulich, A., Cozen, W., Yeager, M., Holford, T. R., Vajdic, C. M., Davis, S., Leaderer, B., Kricker, A., Schenk, M., Zahm, S. H., Chatterjee, N., Wang, S. S. (2009). Genetic variation in caspase genes and risk of non-Hodgkin lymphoma: a pooled analysis of 3 population-based case-control studies. Blood. 114(2), 264–267.

Law, P.J., Timofeeva, M., Fernandez-Rozadilla, C. et al. (2019). Association analyses identify 31 new risk loci for colorectal cancer susceptibility. Nat Commun. 10, 2154.

Le Marchand L. (2009). Genome-wide association studies and colorectal cancer. Surgical oncology clinics of North America. 18(4), 663–668.

Morgan, M. (2021). BiocManager: Access the Bioconductor Project Package Repository. R package version 1.30.16.

Shaik, A. P., Shaik, A. S., Al-Sheikh, Y. A. (2015). Colorectal cancer: A review of the genome-wide association studies in the kingdom of Saudi Arabia. Saudi journal of gastroenterology. 21(3), 123–128.

Simino, J., Sung, Y. J., Kume, R., Schwander, K., Rao, D. C. (2013). Gene-alcohol interactions identify several novel blood pressure loci including a promising locus near SLC16A9. Frontiers in genetics. 4, 277.

Slowikowski, K. (2021). ggrepel: Automatically Position NonOverlapping Text Labels with 'ggplot2'. R package version 0.9.1. https://cran.r-project.org/package=ggrepel

Tanikawa, C., Kamatani, Y., Takahashi, A., Momozawa, Y., Leveque, K., Nagayama, S., Mimori, M., Ishii, H., Inazawa, J., Yasuda, J., Tsuboi, A., Shimizu, A., Sasaki, M., Yamaji, T. et al. (2018). GWAS identifies two novel colorectal cancer loci at 16q24.1 and 20q13.12, Carcinogenesis. 39, 5, 652–660.

Wang, L., Fan, C., Topol, S. E., Topol, E. J., Wang, Q. (2003). Mutation of MEF2A in an inherited disorder with features of coronary artery disease. Science. 302(5650), 1578–1581.

Wickham, H., François, R., Henry, L., Müller, K. (2018). dplyr: A Grammar of Data Manipulation. R package version 0.7.6. https://CRAN.R-project.org/package=dplyr

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. https://cran.r-project.org/package=ggplot2

Zheng, X., Levine, D., Shen, J., Gogarten, S., Laurie, C., Weir, B. (2012). A High-performance Computing Toolset for Relatedness and Principal Component Analysis of SNP Data. Bioinformatics. 28(24), 3326-3328.

SUPPLEMENTARY DATA

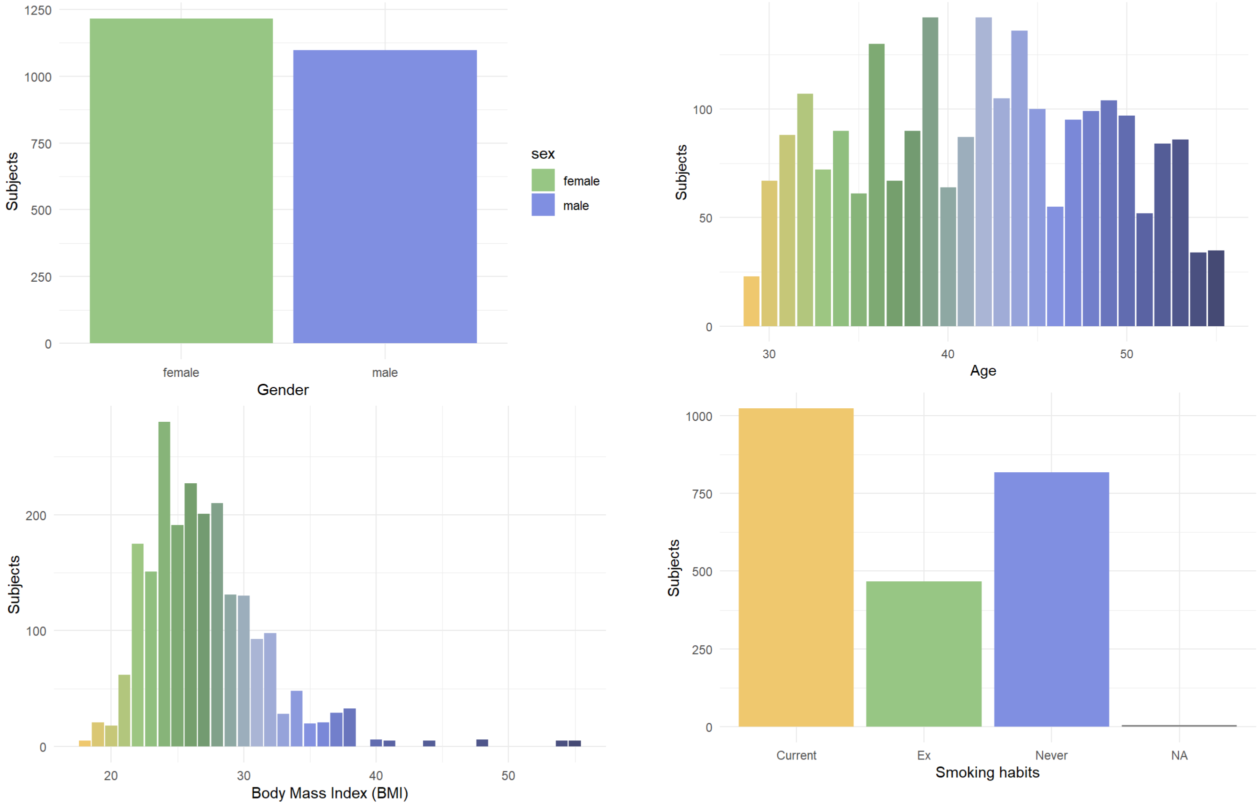


Figure 1. Dataset description distribution according to different parameters: gender, age, body mass index (BMI) and smoking habits.

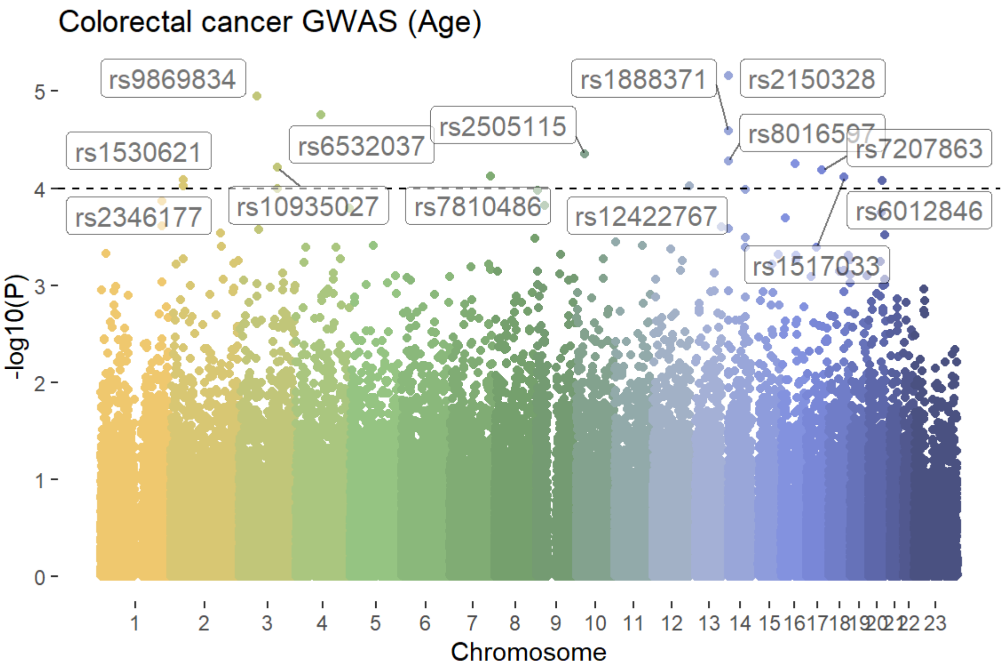


Figure 2. Manhattan plot showing the association p-values for the studied SNPs after adjusting for **age**. The dashed line represents a significance threshold of P= 5 × 10-8. Loci showing a significant association are labeled.

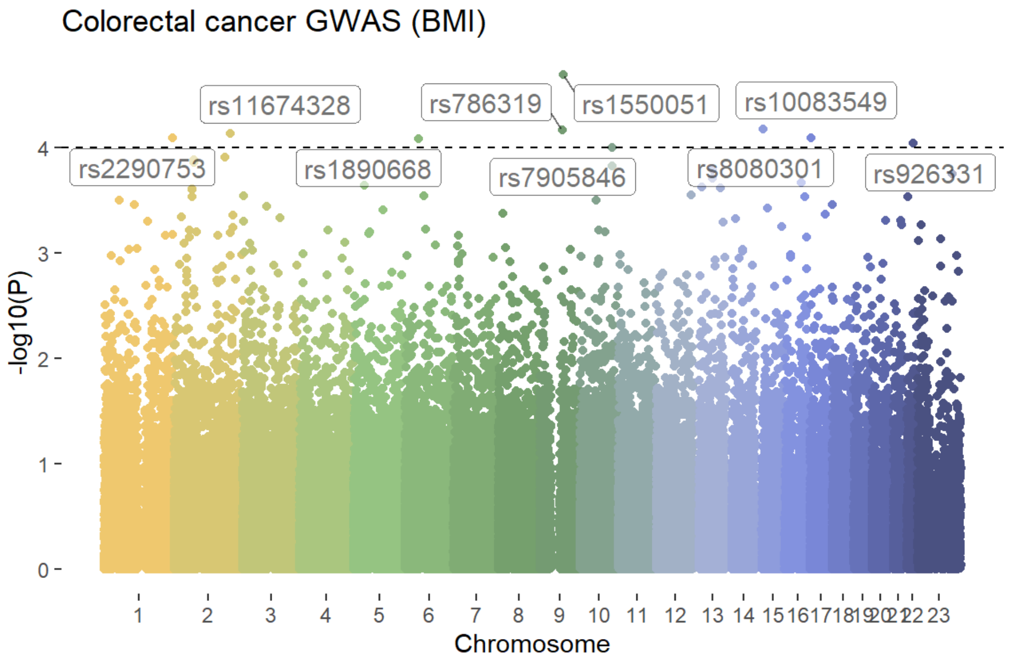


Figure 3. Manhattan plot showing the association p-values for the studied SNPs after adjusting for **body mass index (BMI)**. The dashed line represents a significance threshold of P= 5 × 10-8. Loci showing a significant association are labeled.

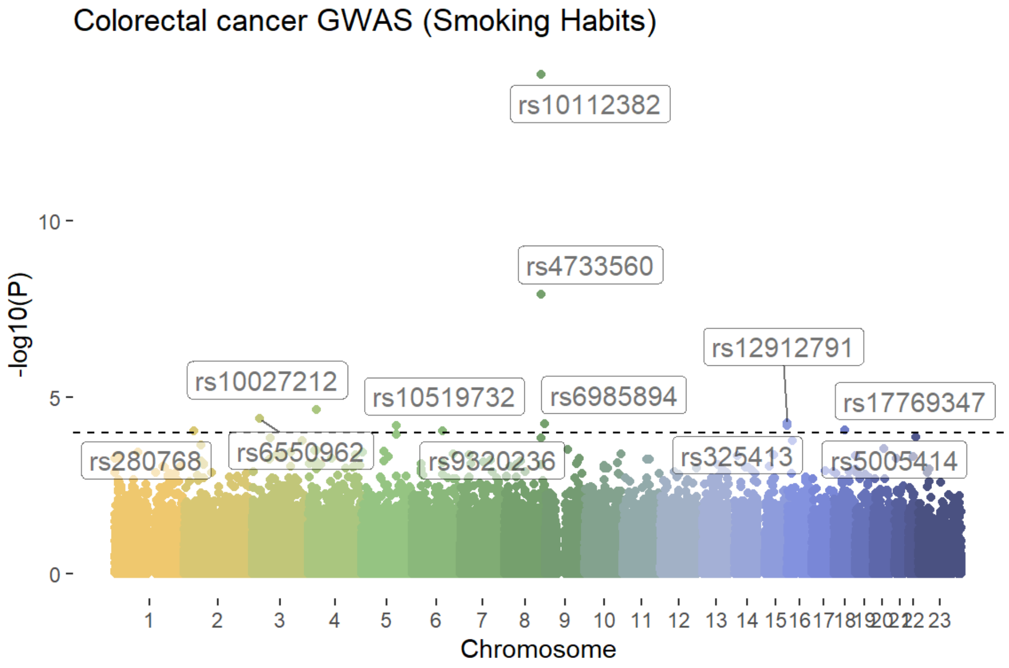
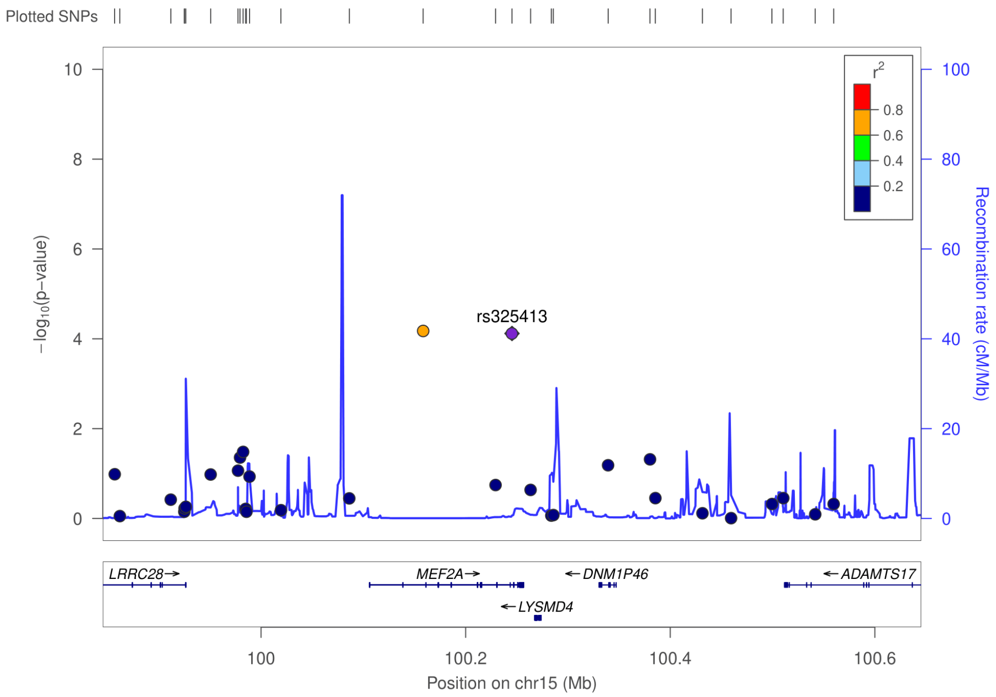
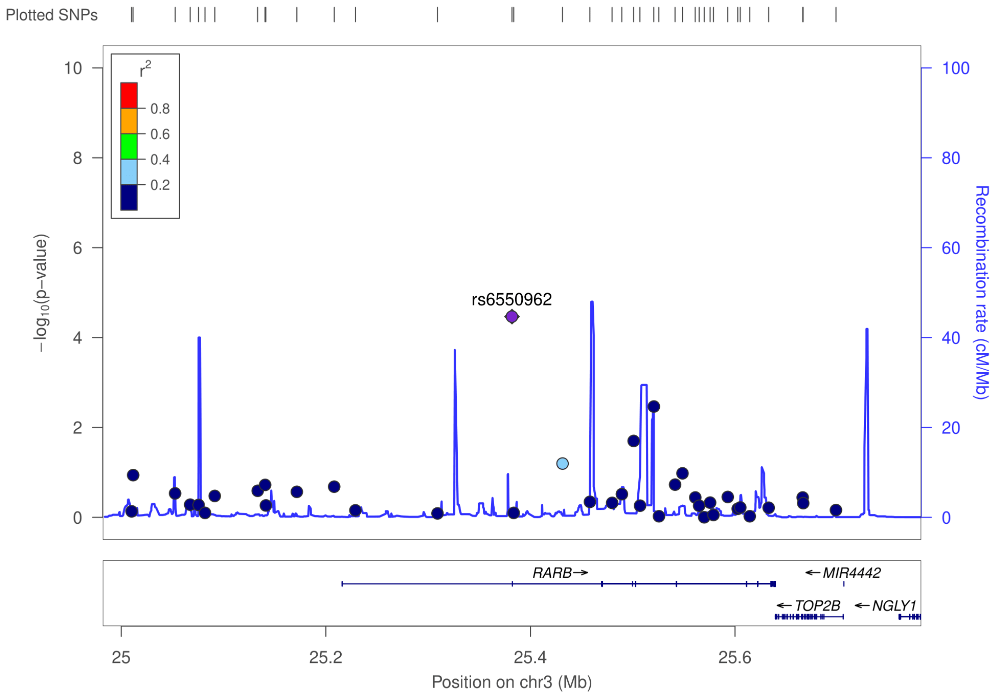


Figure 4. Manhattan plot showing the association p-values for the studied SNPs after adjusting for smoking habits. The dashed line represents a significance threshold of P= 5 × 10-8. Loci showing a significant association are labeled.


Plotted SNPs
10
8
6
4
2
0
−log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
rs325413
LRRC28→
MEF2A→
←DNM1P46
←ADAMTS17
←LYSMD4
100
100.2
100.4
100.6
Position on chr15 (Mb)


Plotted SNPs
10
8
6
4
2
0
−log10(p−value)
100
80
60
40
20
0
Recombination rate (cM/Mb)
r2
0.8
0.6
0.4
0.2
rs6550962
RARB→
←MIR4442
←TOP2B
←NGLY1
25
25.2
25.4
25.6
Position on chr3 (Mb)

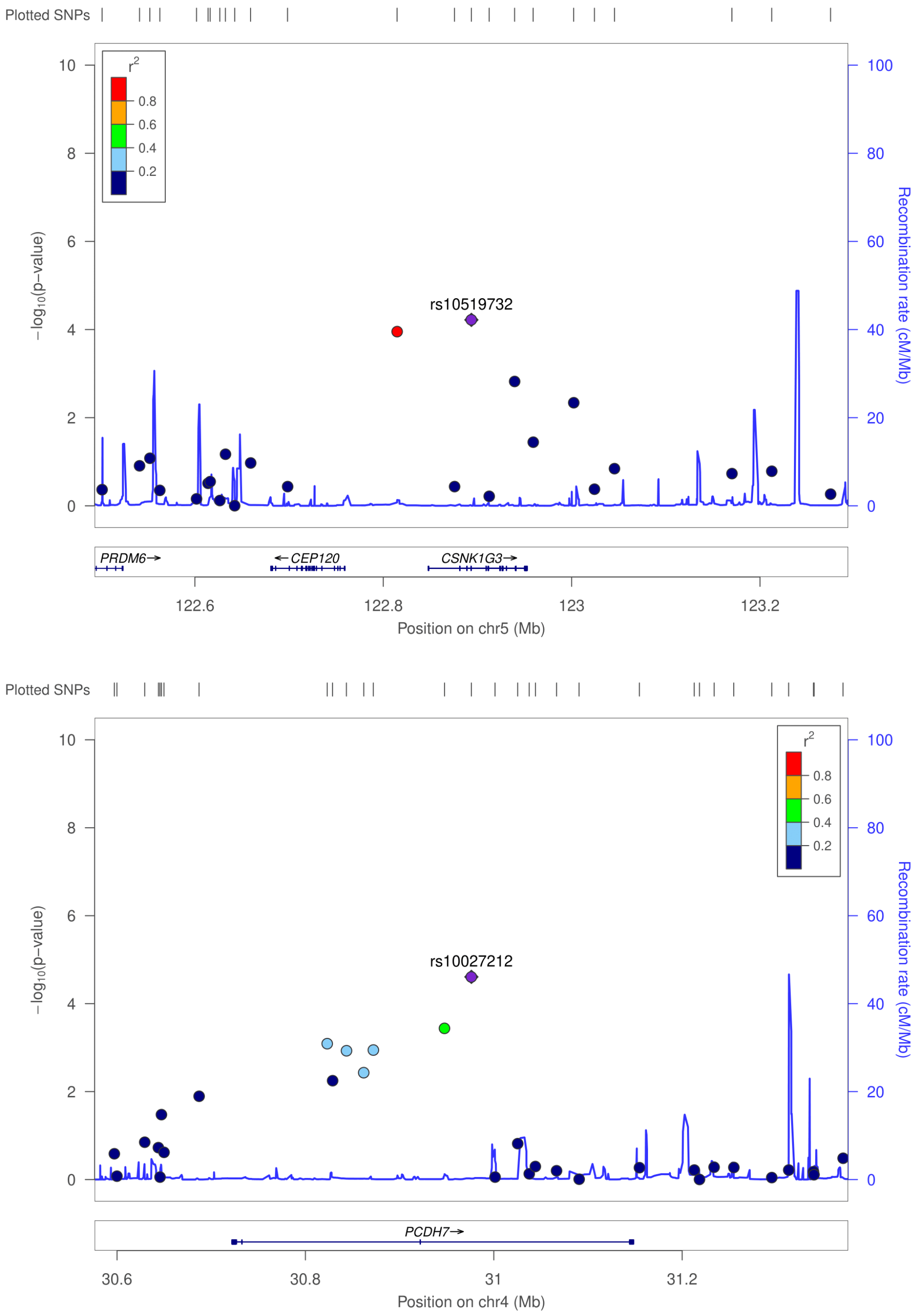


Figure 5. Genomic region representation (400 kb depicted) for SNPs overlapping genes with significant associations.