Temes Actuals de la Bioinformàtica

Pages 1–4
2022

A GWAS study to identify differential SNPs in colorectal cancer

Gisela de Miguel Garcia, Olivia Dove Estrella, Oriol Gómez Lores, Roger Parramon Codina

Universitat Autònoma de Barcelona

January 2022

Keywords: colorectal cancer, GWAS, SNPs, Manhattan plot

ABSTRACT

According to GLOBOCAN 2020 colorectal cancer (CRC) is the second common cancer in women and third most common in men, with an estimated 1.9 million new diagnosed cases worldwide in 2020 [1]. In 2020, about 1 million patients died from CRC, making it the fifth leading cause of cancer-related deaths [1]. Despite strong hereditary components, most cases of CRC are sporadic and develop slowly over several years through the adenoma-carcinoma sequence [2]. There are several factors that increase the risk of CRC, such as age, personal or family history of CRC or adenomatous polyps, and a personal history of chronic inflammatory bowel disease. Also, other factors are diet, micronutrient deficiency, physical activity and obesity, smoking and alcohol consumption [3].
Genome-Wide Association Study (GWAS) has become increasingly used to identify associations between single nucleotide polymorphisms (SNPs) and phenotypic traits, and are commonly applied within the social sciences [4].
In this article we conducted an association study performing a GWAS that includes 100,000 SNPs using real data from a case-control study, with the objective of finding more prevalent SNPs within people with CRC.

METHODS

In this section, we provide the packages and tools used, a data description as well as the quality control performed. Also, we needed the post analytical tools describe below, to find the genes in which the SNPs could be found and to establish their function:

- **LocusZoom:** Maps specific location of each SNP on the genome, showing results in a window of 400kb [5].
- **NCBI:** Used to look for main gene description [6].
- **Gene Ontology:** Used to find each gene's function/s [7].

Packages and tools used

All the analyses were performed using the programming language R version 4.1.2. [8].

- **ggplot2**: R data visualization package.
- **ggrepel**: R package that builds on ggplot2, allowing better data labelling in graphs.
- **dplyr**: R package that provides the necessary grammar for data frames.
- **devtools**: R package necessary to access GitHub and download the SNPassoc package.
- **BiocManager**: R package that enables Bioconductor's software, which helps with the analysis of biological data.
- **SNPassoc**: R package from Bioconductor that performs whole genome association studies, providing tools for descriptive statistics and exploratory analysis of missing values, calculation of Hardy-Weinberg equilibrium, make analysis of association based on generalized linear models (either for quantitative or binary traits), and analysis of multiple SNPs (haplotype and epistasis analysis).
- **snpStats**: R package from Bioconductor used to adjust the analysis data according to clinical, demographic, etc., variables.
- **SNPRelate**: R package from Bioconductor that provides a binary format for single-nucleotide polymorphism (SNP) data in GWAS.

Data description

To reliably identify variation, many subjects are needed. Therefore the GWAS analysis includes 100,000 SNPS, using real data from a case-control study with 2312 genotyped individuals. The phenotype of interest is patients diagnosed with CRC, thus, patients and the controls were differentiated by the value 'cascon' (cascon=0 were controls, cascon=1 were patients with CRC).

Quality Control

Firstly, the control subjects were removed (1,138 individuals) because they were not CRC patients. Then, a quality control (QC) of genomic data (SNPs) was also required before the association testing. The different measures applied in the QC of SNPs are described below:

- **SNPs with high rate of missing**: typically, markers with a call rate less than 95% are removed from association analyses.
- **Rare SNPs (i.e. having low minor allele frequency - MAF)**: markers with a low MAF (¡5%) are usually filtered, too.
- **SNPs that do not pass the Hardy-Weinberg equilibrium (HWE) test**: the significance threshold rejecting a SNPs for not being in HWE varies greatly between studies, but typically a parsimonious threshold of 0.001 may be considered. These values correspond to a z-score of ±3.3. Strictly speaking, HWE test should be applied to controls only.

The total of SNPs removed in the QC were 1,1479: 875 for bad call rate; 10,669 for low MAF and 72 for not having passed the HWE.

Finally, a quality control of individuals was performed. This consisted in the following main steps: identification of individuals with discordant reported and genomic sex, identification of individuals with outlying missing genotype or heterozygosity rate, identification of duplicated or related individuals and identification of individuals of divergent ancestry, from the sample.

Sex discrepancies: Gender is usually inferred from the heterozygosity of chromosome X. Males have an expected heterozygosity of 0 and females

of 0.30. Figure 1 shows that some males were reported with non-zero X-heterozygosity and females with zero X-heterozygosity. These were later identified and excluded from the study.

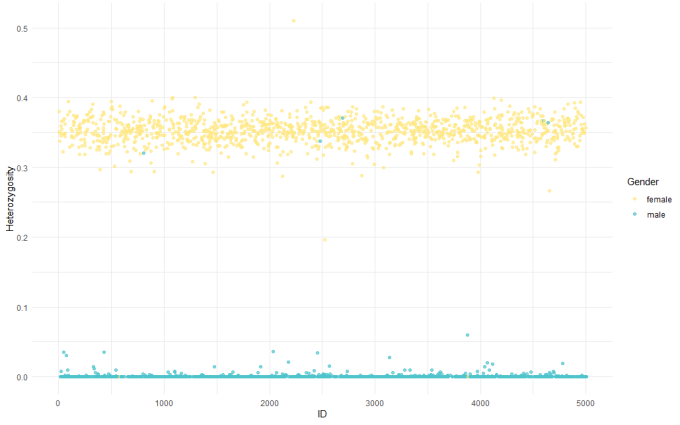

Figure 1. Chromosome X heterozygosity by genders (male/female).

Outlying heterozygosity: We identified individuals with outlying heterozygosity from the overall genomic heterozygosity rate. In figure 2, there are statistical comparisons of these two values per individual, obtaining an F statistic. Individuals whose F statistic is outside the band ±0.1 are considered sample outliers and correspond to those having a heterozygosity rate different from 0.32.

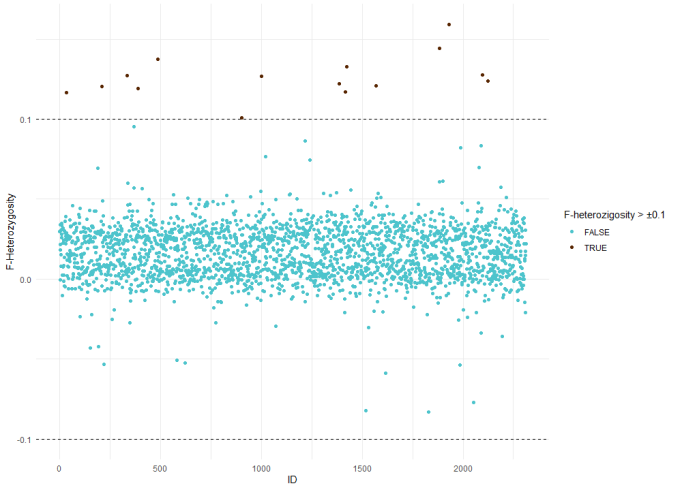

Figure 2. F statistic of each individual after the outlying heterozygosity study.

Close familial relatedness: GWAS are typically studies based on population samples. Therefore, close familial relatedness is not representative of the sample.

Accordingly, individuals whose relatedness was higher than expected doing identity-by-descent (IBD) analysis were identified. The result of such analysis is a table indicating kinship among pairs of individuals. Summing up, individuals with more than 5% missing genotypes, with sex discrepancies, F-heterozygosity absolute value different than ±0.1 and kinship coefficient ¿ 0.1 were removed from the genotype and phenotype data. According to this, from 2,312 individuals we kept 2,243 (Table 1).

Table 1. Summary of the removed individuals by each criteria.

Individuals removed for bad call rate	32
Individuals removed for heterozygosity problems	15
Individuals removed for sex discrepancies	9
Individuals removed due to close familial relatedness	15
Total number of individuals excluded	69

RESULTS

A useful way to summarize genome-wide association data is with a Manhattan plot. This type of plot has a point for every SNP or location tested with the position in the genome along the x-axis and the -log10 p-value on the y-axis. Eleven differential SNPs were identified on the Manhattan plot (Figure 3). Additionally, the location of these SNPs, and a possible link between them with determined genes, were established using LocusZoom.The results of the LocusZoom are summed up on Table 2.

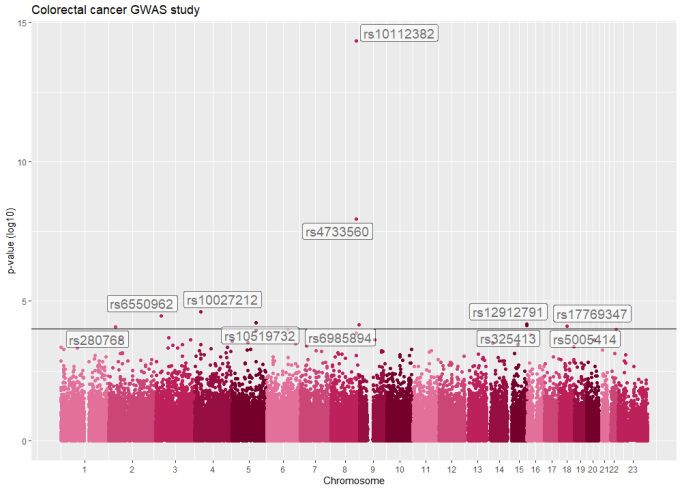

Figure 3. Manhattan plot representation. Significant SNPs are framed.

Table 2. Summary of LocusZoom results. Included are the location and possible genes affected by each SNP.

SNPs id	Location	Genes
rs280768	chr2:35123423	Intergenic region
rs6550962	chr3:25381964	*RARB*
rs10027212	chr4:30976208	*PCDH7*
rs10519732	chr5:122893560	*CSNK1G3*
rs4733560	chr8:128779001	Intergenic region
rs10112382	chr8:128784397	Intergenic region
rs6985894	chr8:143681901	Intergenic region
rs12912791	chr15:100158444	*MEF2A*
rs325413	chr15:100245296	*MEF2A*
rs17769347	chr18:38735059	Intergenic region
rs5005414	chr18:38728795	Intergenic region

DISCUSSION

The GWAS results have yielded 6 SNPs corresponding to intergenic regions (rs280768, rs4733560, rs10112382, rs6985894, rs17769347,

rs5005414) and 5 SNPs falling into genic regions (rs6550962, rs10027212, rs10519732, rs12912791, rs325413). All genes encode for proteins, and according to their function:

- ***RARB* gene:** This gene is linked to cell/tissue differentiation because it is a nuclear transcriptional regulator. *RARB* has been seen to partake in the digestive tract development at embryonic stages. Mutations in this gene may be associated with a loss of function, thus, possibly leading to cellular dedifferentiation.
- ***PCDH7* gene:** This gene is linked to cell adhesion because it is an integral membrane protein. Mutations in this gene may be associated with weakening of cell union, possibly promoting a metastatic behaviour.
- ***CSNK1G3* gene:** This gene is linked to PTM addition, and therefore, signal transduction. Mutations in this gene may affect cellular pathways (Wnt pathway), like those related to differentiation. If these pathways are mutated, there may be dedifferentiation and larger proliferation.
- ***MEF2A* gene:** This gene is associated with muscular tissue differentiation, acting as a co-activator with a master regulator gene (Myo-D) to activate muscular-specific genes. Furthermore, this gene controls the inactivation of pluripotent genes. Mutations in this gene may affect their silencing and favour cellular dedifferentiation and proliferation. Moreover, two SNPs in this gene have been found as a result of the GWAS analysis to be associated with CRC. This suggests that *MEF2A* gene may be relevant for the onset of this cancer.

Intergenic regions possibly play an interesting role. In this study, LocusZoom has spotted SNPs belonging to intergenic regions either surrounded by genes or stranded within the 400kb window. Irrespective of the intergenic SNPs location, it is possible that these SNPs may be cis-regulatory elements, acting as enhancers for proximal genes or more distanced genes due to the chromatin 3D structure (because of TADs, for example). With a GWAS study, these hypotheses can not be confirmed. Options to confirm this include chromatin structure analysis like Hi-C, followed by different experiments which analyze gene expression and promoters activation. This may lead to the discovery of more genes implicated in cancer initiation/development.

This is a GWAS study. It must be remembered that a SNP found with a GWAS analysis might not be associated with the illness, but it may be in linkage disequilibrium (LD) with the actual gene responsible for the illness. Therefore, GWAS results should be treated with caution. It is difficult to choose which SNPs are good candidates for molecular study. Consequently, it is advisable to check further sources and experiments by other laboratories in order to check whether the SNPs, this GWAS study has identified, should be investigated further. Should we only consider our GWAS results, SNPs which fall into intergenic regions may be less important than those which fall into a protein-coding gene locus. Within this latter group, the most interesting SNPs to study are *MEF2A*'s ones. This is because the GWAS study has found two SNPs falling within this gene, while there is just one SNP for the other genes. Lastly, cancer is a multifactorial disease but only genetic traits are studied in GWAS. The etymology of cancer is not only genetic but also environmental, in order to establish reliable causality for CRC, environmental studies should also be carried out.

REFERENCES

[1] Globocan 2020. Colorectal cancer. Retreived from: https://gco.iarc.fr/, 2020. Accessed: 2021-12-30.

[2] Hermann Brenner, Matthias Kloor, and Christian Peter Pox. Colorectal cancer. *The Lancet*, 383(9927):1490–1502, 2014.

[3] Sabha Rasool, Showkat Ahmad Kadla, Vamiq Rasool, and Bashir Ahmad Ganai. A comparative overview of general risk factors associated with the incidence of colorectal cancer. *Tumor Biology*, 34(5):2469–2476, 2013.

[4] Andries T. Marees, Hilde de Kluiver, Sven Stringer, Florence Vorspan, Emmanuel Curis, Cynthia Marie-Claire, and Eske M. Derks. A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608, 2018. e1608 IJMPR-Feb-2017-0014.R3.

[5] Andrew P Boughton, Ryan P Welch, Matthew Flickinger, Peter VandeHaar, Daniel Taliun, Gonçalo R Abecasis, and Michael Boehnke. Locuszoom.js: Interactive and embeddable visualization of genetic association study results. *Bioinformatics*, 37(18):3017–3018, 2021.

[6] Richa Agarwala, Tanya Barrett, Jeff Beck, Dennis A Benson, Colleen Bollin, Evan Bolton, Devon Bourexis, J Rodney Brister, Stephen H Bryant, Kathi Canese, and et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 46(D1), 2017.

[7] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, and et al. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000.

[8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021.