UNIVERSIDAD DE GRANADA

TRABAJO FIN DE MÁSTER

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS E INGENIERÍA DE COMPUTADORES

Characterization of Genetically Different Groups of Cancer Patients Using Unsupervised Methods

Autora: Lucía Almorox Antón.

Tutora: María Coral del Val Muñoz.

Mentora: Elisa Díaz de la Guardia Bolívar.

Convocatoria: Convocatoria Especial de Fin de Estudios, diciembre 2023.

Fecha: 10 de diciembre de 2023.

Yo, Lucía Almorox Antón, alumna del Máster Universitario en Ciencia de Datos e Ingeniería de Computadores de la Escuela Internacional de Posgrado de la Universidad de Granada, con DNI 51550791V, declaro de manera explícita y firmo la presente declaración, asumiendo la completa originalidad del trabajo titulado "Characterization of Genetically Different Groups of Cancer Patients Using Unsupervised Methods". Confirmo que en la elaboración de este trabajo no he utilizado fuentes sin citarlas debidamente.

ALMOROX ANTON LUCIA - 51550791V

Firmado digitalmente por ALMOROX ANTON LUCIA - 51550791V
Fecha: 2023.12.10 15:53:09 +01'00'

Fdo: Lucía Almorox Antón.

Characterization of Genetically Different Groups of Cancer Patients Using Unsupervised Methods.

Lucía Almorox Antón

December 10, 2023

General index

Index of Figures v

Index of Tables viii

List of Abbreviations x

Abstract xi

Resumen xii

Key Words xiii

CHAPTER 1: INTRODUCTION AND OBJECTIVES 1

1.1 Foundations of Genetics: Understanding DNA, Chromosomes, and Gene Expression. 1

1.2 Genomics and Single Nucleotide Polymorphisms. 2

1.3 Alleles and Genotypes of SNPs. 3

1.4 Data Science Empowering Precision Medicine. 5

1.5 GWAS and PGMRA. 5

1.6 PGMRA Biclusters. 6

1.6.1 Non-Negative Matrix Factorization (NMF). 6

1.6.2 Bio Non-Smooth NMF (bioNMF). 8

1.6.3 Fuzzy NMF Biclustering Method (FNMF). 8

1.6.4 Identification of Good Clusters without Determining a priori a Specific Number of Maximum Clusters. . . . 9
1.7 Colorectal Cancer. . . . 10
1.7.1 CRC Heritability. . . . 10
1.7.2 Molecular Pathological Classifications for CRC. . . . 10
1.7.3 Risk Factors. . . . 12
1.7.4 Metastasis in CRC. . . . 13
1.8 Project's Motivation and Objectives. . . . 13

CHAPTER 2: INPUT DATA AND METHODOLOGY 16
2.1 Data Description and Methodology Flowchart. . . . 16
2.2 Data Exploration and Preprocessing. . . . 19
2.2.1 Detection of Erroneous Patient Samples and Subsequent Filtering of Patients, Biclusters, and SNPs. . . . 19
2.2.2 Missing Values. . . . 19
2.2.3 Data Engineering on the Clinical Dataset and Visualization of Attributes Distribution. . . . 20
2.2.4 Preprocessing of the Ensembl Dataset and Visualization of Attributes Distribution: Distinguishing Genic and Intergenic SNPs. . . . 20
2.2.5 Exploring Hemizygosity Encoding. . . . 21
2.3 Analysis of Similarity Among Biclusters Regarding SNP and Patient Composition. 22
2.3.1 Analysis of Patient Sharing Among PGMRA Biclusters. . . . 22
2.3.2 Analysis of SNP Sharing Among PGMRA Biclusters. . . . 23
2.4 Extraction and Filtering of Association Rules in the Clinical Dataset. . . . 23
2.5 Identification of Interesting Intergenic SNPs Using RegulomeDB. . . . 25
2.6 Gene Annotation. . . . 26
2.6.1 Search for Associations with Colon Cancer. . . . 26
2.6.2 Search for Associations with Cancer in General. . . . 28
2.6.3 Manual Search for Associations with Colon Cancer or Cancer in General for Genes that VarElect did not Link to Either of the two Phenotypes. . . . 28
2.7 Gene Set Over-Representation Analysis. . . . 29
2.8 Gene Expression Analysis Using External RNA-Seq Data. . . . 30

2.9 Minimum Protein-Protein Interaction Network Derived from All Affected Genes. 31
2.10 Representation of SNPs in Gene Maps. 33
2.11 Bicluster-Specific Analyses. 34
2.12 Identification of Interesting Biclusters and Genes. 35
2.13 Genotypic Cluster-Heatmap. 36

CHAPTER 3: RESULTS AND DISCUSSION 37
3.1 Missing Values. 37
3.2 Hemizygosity Encoding. 37
3.3 Analysis of Similarity Among PGMRA Biclusters. 39
3.3.1 Patient Sharing Among PGMRA Biclusters. 39
3.3.2 SNP Sharing Among PGMRA Biclusters. 41
3.4 Analysis of the Entire Sets of Patients and SNPs. 43
3.4.1 Patient Clinical Data Analysis: Bar Charts and Association Rules. 44
3.4.2 Genic SNPs Analysis. 48
3.4.3 Intergenic SNPs Analysis. 64
3.5 Bicluster-Specific Analyses. 67
3.5.1 Biclusters Overview: Patient, SNP, and Gene Data. 67
3.5.2 Bicluster 16.12. 70
3.5.3 Bicluster 17.8. 81
3.5.4 General Protein-Protein Interaction Network Created from Genes Affected in either Bicluster 16.12 or 17.8. 89
3.6 Identification of Interesting Genes. 90
3.6.1 Genes Affected in Many Biclusters or with Many Associated SNPs. 90
3.6.2 PTCHD1-AS and ADAMTS9-AS1 Genes. 91
3.6.3 AFF2 Gene. 93
3.7 Genotypic Cluster-Heatmap of Biclusters 2.2 and 8.4. 96

CHAPTER 4: CONCLUSIONS AND FUTURE WORK 99

CHAPTER 5. BIBLIOGRAPHY 102

A Introduction - Supplementary Material. 115

A.1 Examples of how any Type of SNP Can Give Rise to a Observable Phenotype. . 115

B Methodology - Supplementary Material. 115

B.1 Illumina Microarray Technology. 115

B.2 Data Engineering Steps on the Clinical Dataset. 116

B.3 Explanations of Some Terms Related to Graph Theory. 116

B.4 Extraction of Association Rules: Brute-force Extraction, Two-step Approach and Apriori algorithm. 117

B.5 RelomeDB Supporting Evidence Ranking. 118

B.6 Filling in the Table of Genes Directly Related to CRC. 118

B.7 Management of Gene Aliases to Identify Genes from Our Dataset Present in the Set of Extracted DEGs. 119

B.8 Louvain Algorithm for Network Community Detection. 119

C Results and Discussion - Supplementary Material. 120

C.1 Distribution of Genic SNP Severity Based on the Chromosome. 120

C.2 Protein-Protein Interaction Network Derived from the Total Set of Affected Genes. 121

C.2.1 Topological and Centrality Analysis. 121

C.2.2 Key Over-represented Terms in the Nodes. 123

C.3 Bicluster-Specific Clinical Data. 124

C.3.1 Bicluster 16.12 Clinical Data. 124

C.3.2 Bicluster 17.8 Clinical Data. 126

C.4 Interesting Genes. 127

C.4.1 Interesting Genes Directly Related to CRC. 127

C.4.2 Interesting Genes Indirectly Related to CRC. 130

C.5 Brief Selection of Genes that Present Haplotypes in the Genotypic Cluster-Heatmap of Biclusters 2.2 and 8.4. 134

Index of Figures

1 Organization of DNA into chromosomes. 1

2 Central dogma of molecular biology: replication, transcription, splicing (considered part of transcription), and translation. 2

3 SNP concept. 3

4 Approximate NMF. 7

5 Biological differences in the gene expression-based molecular subtypes of CRC . . 12

6 Data for analysis. 18

7 Methodology. 18

8 Percentage matching matrix of patients between biclusters. 40

9 Bicluster patients directed weighted graph. 41

10 Percentage matching matrix of SNPs between biclusters. 42

11 Bicluster SNPs directed weighted graph. 43

12 Bar chart representing the number of patients in each bicluster. 44

13 Bar charts of several clinical attributes: "Age_Range", "Sex", "Metastasis", "Surgery", "Histog", "Loc_T", Loc_T2", "Stage", "Degree" and "Basal_CEA". 45

14 Bar charts of several clinical attributes: "Met_liver","Met_lung" and "Met_lymph_nodes". 46

15 Bar charts of key attributes in the genic SNP Ensembl table: "chr_name", "ensembl_type", "consequence_type_tv", "Impact" and "sift_prediction". 49

16 Bar chart of the 15 most over-represented diseases (with lowest adjusted p-value) in the total of affected genes. 52

17 Bar chart of the 15 most over-represented diseases (with lowest adjusted p-value) in the genes for which a direct relationship with CRC has been found. 54

18 50 out of the 72 genes identified as DEGs (from the total of affected genes) in colorectal cancer vs. healthy colon/rectum: boxplots of their expression in the two sample types, after preprocessing of TCGA data. 57

19 72 genes (from the total of affected genes) identified as DEGs in colorectal cancer vs. healthy colon/rectum: heatmap of their expression in the two sample types, after preprocessing of TCGA data. 58

20 Minimum protein-protein interaction network derived from the total set of affected genes. 60

21 Minimum protein-protein interaction network derived from the total set of affected genes. Each node color indicates membership in a different Louvain community. . 62
22 Minimum protein-protein interaction network derived from the total set of affected genes: Circle Pack layout based on community membership. 64
23 Bar chart depicting the Ensembl "chr_name" attribute in the intergenic SNP Ensembl table. 65
24 Summary table of each bicluster (patient, SNP, and gene-related information). . 68
25 Distribution of the variable "Age_Range" based on membership or non-membership to bicluster 16.12. 71
26 Distribution of the variable "Metastasic" based on membership or non-membership to bicluster 16.12. 72
27 Summary table of genic SNPs in bicluster 16.12. 74
28 Bar chart of the over-represented diseases in the genes affected in bicluster 16.12. 75
29 Degree 1 general protein-protein interaction network created from the genes affected in bicluster 16.12. 81
30 Distribution of the variable "Age_Range" based on membership or non-membership to bicluster 17.8. 82
31 Distribution of the variable "Metastasic" based on membership or non-membership to bicluster 17.8. 83
32 Summary table of filtered intergenic SNPs in bicluster 17.8. 84
33 Summary table of genic SNPs in bicluster 17.8. 85
34 Degree 1 general protein-protein interaction network created from the genes affected in bicluster 17.8. 88
35 Minimum general protein-protein interaction network of degree 1 created from the genes affected in biclusters 16.12 and 17.8. 89
36 ADAMTS9-AS1 gene map. 92
37 Gene expression regulation exerted by PTCHD1-AS on PTCHD1, which could be influenced in CRC by PTCHD1-AS SNPs. 93
38 PTCHD1-AS gene map of bicluster 2.2. 93
39 AFF2 gene map of bicluster 14.4. 94

40 Segment of the summary table of genic SNPs in bicluster 14.4 displaying the SNPs located in the AFF2 gene. 95
41 Portion of a long-format genotypic heatmap for biclusters 2.2 and 8.4. 97
42 RegulomeDB scoring scheme meaning. 118
43 Genic SNP Ensembl table: Distribution of the variable "consequence_type_tv" based on the variable "chr_name" 120
44 Genic SNP Ensembl table: Distribution of the variable "Impact" based on the variable "chr_name" 121
45 Minimum protein-protein interaction network derived from the total set of affected genes. The circular layout is used to show degree 1 seed nodes conected to UBC. 122
46 Bar charts of several clinical attributes in bicluster 16.12: "Age_Range", "Sex", "Metastasis", "Surgery", "Histog", "Loc_T", Loc_T2", "Stage", "Degree", "Basal_CEA".124
47 Bar charts of several clinical attributes in bicluster 16.12: "Met_liver","Met_lung", "Met_lymph_nodes". 125
48 Bar charts of several clinical attributes in bicluster 17.8: "Age_Range", "Sex", "Metastasis", "Surgery", "Histog", "Loc_T", Loc_T2", "Stage", "Degree" and "Basal_CEA". 126
49 Bar charts of several clinical attributes in bicluster 17.8: "Met_liver", "Met_lung" and "Met_lymph_nodes". 127
50 TNFRSF17 gene map. 128
51 GSDMB gene map of bicluster 2.2. 129
52 DMD gene map of bicluster 2.2. 130
53 PCDH11X gene map of bicluster 11.5. 131
54 GSTM3 gene map. 132
55 GSTM5 gene map. 132
56 GRIA3 gene map of bicluster 3.2. 133

Index of Tables

1 TCGA Data: Downloaded, randomly selected (undersampling) and filtered samples of each class. 31

2 Subset of the association rules that were generated with a length between 2 and 4 items, a minimum support threshold of 0.05, a minimum confidence threshold of 0.7, and a minimum lift threshold of 1.4. 47

3 Number of genes for which a relationship (direct or indirect) with CRC or cancer in general has been found. 51

4 Over-represented Cellular Component GO terms in the total set of affected genes. 53

5 Over-represented Cellular Component GO terms in the set of genes directly related to CRC. 54

6 Over-represented pathways in the set of genes directly related to CRC. 56

7 Percentage of genes for which a relationship (direct or indirect) with CRC or cancer in general has been found: a comparison in the total set of affected genes and those identified as DEGs. 59

8 Selection of over-represented terms (according to Cytoscape) in each Louvain community of genes. The p-value for each term is displayed next to it in parentheses. 63

9 Intergenic SNPs with a probability of 0.25 or higher of having regulatory function in the colon, large intestine or intestine according to RegulomeDB. 66

10 Over-represented MF GO terms in bicluster 16.12 genes. 75

11 Over-represented pathways in bicluster 16.12 genes. 76

12 Genes affected in bicluster 16.12: description, relationship with actin or muscle, additional information, and biclusters in which each gene appears. The source of the data in columns 2-4 is GeneCards, unless otherwise specified in the cell. . . . 77

13 Over-represented MF GO terms in bicluster 17.8 genes. 86

14 Genes affected in bicluster 17.8: description, additional information, and biclusters in which each gene appears. The source of the data in columns 2-3 is GeneCards. 87

15 Top 5 ranking of genes affected in the highest number of biclusters. 90

16 Top 5 ranking of genes affected in the highest number of biclusters 91

17 Key over-represented terms in the nodes of the minimum protein-protein interaction network constructed from the total of affected genes. Information extracted using NetworkAnalyst. 123
18 Some of the genes that seem to present haplotypes in the genotypic cluster-heatmap of biclusters 2.2 and 8.4. 134

List of Abbreviations

CRC	Colorectal Cancer
DNA	Deoxyribonucleic Acid
RNA	Ribonucleic Acid
mRNA	Messenger Ribonucleic Acid
SNP	Single Nucleotide Polymorphism
MSI	Microsatellite Instability
IBD	Inflamatory Bowel Disease
GWAS	Genome-Wide Association Study
PGMRA	Phenotype x Genotype Many-to-many Relation Analysis
NMF	Non-Negative Matrix Factorization
FNMF	Fuzzy Non-Negative Matrix Factorization
DEG	Differentially Expressed Gene
ORA	Over-Representation Analysis
GO	Gene Ontology
CC	Cellular Component
MF	Molecular Function
BP	Biological Process
eQTL	Expression Quantitative Trait Loci
EMT	Epithelial-mesenchymal transition
RBP	RNA-Binding Protein
LD	Linkage Disequilibrium

Abstract

Patients' genetic variability can significantly influence manifestations of colorectal cancer (CRC), resulting in a great heterogeneity of clinical symptoms. This project proposes to explore the performance of the PGMRA unsupervised algorithm on small datasets and evaluate its capacity to generate new knowledge. PGMRA was able to identify fuzzy biclusters (CRC patients x SNPs) based on genotypic data from patients with different diagnoses. The clinical profiles of the patients and the impact of the SNPs included in the CRC biclusters were extensively analyzed, utilizing a broad range of analytical and bioinformatic techniques. The analyses were performed on all patients together and compared to the biclusters' results.

The analysis of the entire sets of patients and SNPs uncovered the functional annotation of 695 affected genes, establishing direct relationships with CRC for 146 of them. Over-representation analyses identified terms related to neurological disorders in the total gene set and additional terms linked to calcium regulation in genes directly related to CRC. Both neurological disorders and hypercalcemia are known to be highly prevalent conditions in CRC. These prevalences could be originated by some of the genic SNPs that PGMRA utilized to group CRC patients.

PGMRA biclusters arose interesting findings. Bicluster 16.12 included a small number of patients and genic SNPs, characterized by generally high MAFs (minor allele frequencies). Notably, this bicluster showed the lowest average age (44.25). The observed gene functions, related to inflammation and actin reorganization, point to classify its members with the CRC Consensus Molecular Subtype 4 (CMS4 - Mesenchymal). Additionally, these same CRC cases could be associated to Inflammatory Bowel Disease, explaining the low average age.

We identified that AFF2 and PTCHD1-AS genes appeared frequently in different biclusters, affected by a high number of SNPs and remarkably high MAFs in some cases. These combinations varied across biclusters. Moreover, our RNA-Seq analysis, using external samples, identified both AFF2 and the target gene of PTCHD1-AS as underregulated genes in CRC. Thus, PGMRA results on small sample sizes showed to be able to find new insights into the complex interplay between genetic variants, clinical characteristics, and potential new perspectives on CRC subtypes, that must be experimentally tested in the laboratory.

Resumen

La variabilidad genética de los pacientes puede influir significativamente en las manifestaciones del cáncer colorrectal (CRC), resultando en una gran heterogeneidad de síntomas clínicos. Este proyecto propone explorar el rendimiento del algoritmo no supervisado PGMRA en conjuntos de datos pequeños y evaluar su capacidad para generar nuevo conocimiento. PGMRA fue capaz de identificar biclústeres difusos de CRC (pacientes de CRC x SNPs) a partir de datos genotípicos de pacientes con diferentes diagnósticos. Los perfiles clínicos de los pacientes y el impacto de los SNPs incluidos en los biclústeres de CRC fueron extensamente analizados, utilizando diversas técnicas analíticas y bioinformáticas. Los análisis se realizaron sobre el conjunto completo de pacientes y se compararon con los resultados obtenidos en los distintos biclústeres.

El análisis del conjunto completo de pacientes y SNPs reveló la anotación funcional de 695 genes afectados, estableciendo relaciones directas con el CRC para 146 de ellos. Los análisis de sobre-representación identificaron términos relacionados con trastornos neurológicos en el conjunto total de genes y términos adicionales vinculados a la regulación del calcio en genes directamente relacionados con el CRC. Se sabe que los trastornos neurológicos y la hipercalcemia son condiciones altamente prevalentes en el CRC. Estas prevalencias podrían estar originadas por algunos de los SNPs génicos que PGMRA utilizó para agrupar a los pacientes de CRC.

Los biclústeres de PGMRA revelaron resultados interesantes. El biclúster 16.12 incluyó un número pequeño de pacientes y SNPs génicos, caracterizados por frecuencias del alelo menor (MAFs) generalmente altas. Notablemente, este bicluster presentó la edad media más baja (44.25). Las funciones génicas observadas, relacionadas con la inflamación y la reorganización de la actina, apuntan a clasificar a sus miembros con el Subtipo de CRC Molecular de Consenso 4 (CMS4 - Mesenquimal). Adicionalmente, estos mismos casos de CRC podrían estar asociados a enfermedades inflamatorias intestinales, lo que explicaría la baja edad media.

Identificamos que los genes AFF2 y PTCHD1-AS aparecían frecuentemente en diferentes biclústeres, afectados por un alto número de SNPs y con MAFs especialmente altas en algunos casos. Estas combinaciones variaban en los biclústeres. Además, nuestro análisis de RNA-Seq, utilizando muestras externas, identificó tanto a AFF2 como al gen objetivo de PTCHD1-AS como genes subregulados en el CRC. En definitiva, los resultados de PGMRA en tamaños de muestra pequeños

mostraron habilidad para encontrar nuevas perspectivas sobre la compleja interacción entre variantes genéticas, características clínicas y posibles nuevos enfoques en los subtipos de CRC, que deben ser probados experimentalmente en el laboratorio.

Key Words

Genotype, Single Nucleotide Polymorphism (SNP), Gene, Colorectal Cancer (CRC), Unsupervised Learning, PGMRA, Bicluster.

CHAPTER 1: INTRODUCTION AND OBJECTIVES

1.1 Foundations of Genetics: Understanding DNA, Chromosomes, and Gene Expression.

DNA (Deoxyribonucleic Acid) is a molecule essential for the growth, functioning, and inheritance of all living organisms. It is composed of four **nucleotides** —adenine (A), thymine (T), cytosine (C), and guanine (G)— arranged in unique sequences that carry the genetic information [1]. As shown in Figure 1, in humans, nuclear DNA is organized into thread-like structures called **chromosomes**. Each chromosome houses **genes**, that can be interpreted, in a simplified manner, as chapters encoding instructions for various biological functions [2].

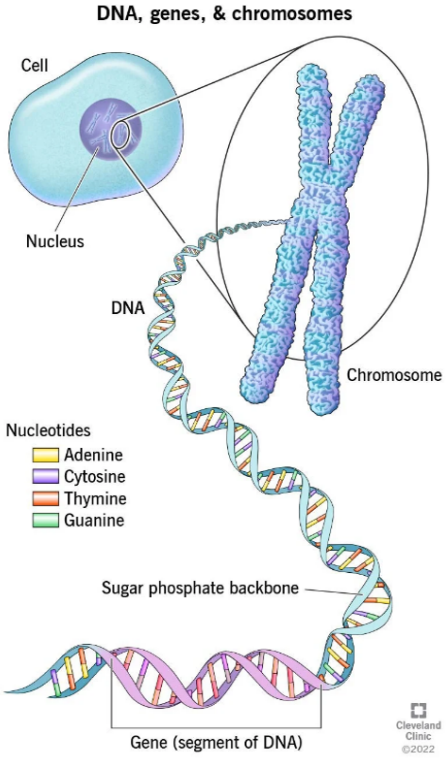


Figure 1. Organization of DNA into chromosomes. Figure taken from a external source [3].

Within this genetic landscape, a fundamental principle known as the **Central Dogma of Molecular Biology** comes to light. This principle, simplified in Figure 2, elucidates the flow of genetic information within a cell. Most genes are initially transcribed into **pre-mRNA**, a precursor molecule that undergoes a crucial process known as **splicing**, in which specific sequences, known as **introns**, are removed. This results in the formation of mature messenger RNA (**mRNA**). Sub-

sequently, this mature mRNA serves as a template for the **protein synthesis** (translation).[4, 5]

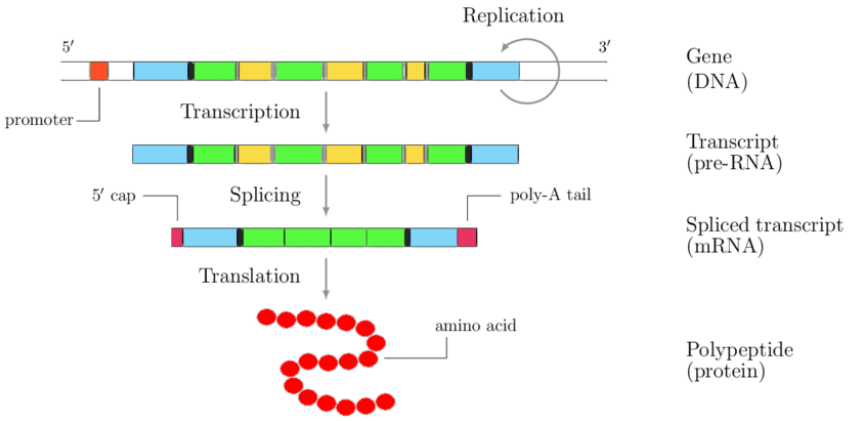


Figure 2. Central dogma of molecular biology: replication, transcription, splicing (considered part of transcription), and translation. In this diagram, green-colored regions represent exons, and yellow regions represent introns. Figure taken from an external source [6].

Between genes are **intergenic regions**, stretches of DNA not involved in protein coding, yet potentially holding crucial regulatory roles for gene control and activity [7].

1.2 Genomics and Single Nucleotide Polymorphisms.

A **genome** is the complete set of DNA (genetic material) present in an organism and the field that delves into the study of entire genomes is called **genomics** [8]. Human genomes contain slight variations called **single nucleotide polymorphisms (SNPs)** [9] [1], as depicted in Figure 3. One of these variations must occur in at least 1% of the population to be regarded as a SNP; otherwise, it would be classified as a point mutation [10].

[1] In biology, the term polymorphism describes the existence of multiple forms, also known as morphs.

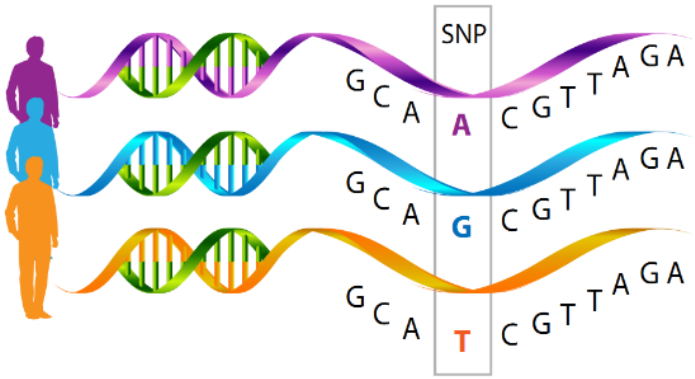


Figure 3. SNP concept. Figure taken from a external source [11].

SNPs contribute to diversity among individuals and, in some cases, can predispose one to disease or influence the response to substances or agents. [8] SNPs may change the encoded amino acids (non-synonymous) or can be silent (synonymous) or simply occur in the non-coding regions. They may influence promoter activity (gene expression), mRNA conformation (stability), and subcellular localization of mRNAs and/or proteins [12]. Some examples of how any type of SNP can give rise to a disease or observable phenotype can be found in Appendix A.1.

1.3 Alleles and Genotypes of SNPs.

An **allele** or allelomorph is each of the alternative forms that a sequence of nucleotides can take at a particular locus [2] on a DNA molecule. Humans, as **diploid** organisms, inherit two sets of chromosomes – one from each parent. This generally results in each locus having **two alleles** in the same individual [13].

A SNP can have, at least, two possible **alleles** in a population:

- **Major allele (A)**: The most common allele in a population. It represents the most prevalent nucleotide at a specific genomic position, within a given reference genome.
- **Minor allele (B)**: The least common allele in a population. It represents the less frequent nucleotide in that genomic position.

Minor allele frequency (MAF) is the fraction of all chromosomes in the population that carry the minor allele over the total population [14, 15].

[2] A "locus" (plural: "loci") in genetics refers to the specific physical location or position of a gene, SNP, or other significant sequence on a chromosome.

In the context of a disease, the term **"risk allele"** denotes the allele that imparts a susceptibility to developing the disease. Generally, the risk allele coincides with the minor allele, as the majority of individuals do not carry the risk allele.

The **genotype of a SNP** refers to the specific combination of alleles an individual has for a particular SNP. The three main genotypes for a SNP are:

- **Homozygous Major (AA)**: an individual has two copies of the major allele.
- **Heterozygous (AB)**: an individual has one copy of the major allele and one copy of the minor allele.
- **Homozygous Minor (BB)**: an individual has two copies of the minor allele.

A special case of genotypes emerges in the concept of hemizygosity, which represents a unique genetic state. Chromosomes can be autosomal (pairs 1-22) or sex chromosomes (pair 23). **Sex chromosomes** (X and Y) largely govern the distinctions between males (XY) and females (XX). Most loci on the X chromosome do not have a counterpart on the Y chromosome, and vice versa. Therefore, since males carry only one X and one Y chromosome, they are unable to be homozygous or heterozygous for alleles exclusive to either of the sex chromosomes. Instead, the term "hemizygous" denotes their allelic condition [16] [3].

However, it is important to consider that in humans, there are two shared regions of homology between the sex chromosomes (**pseudoautosomal regions**). This implies that loci located in these regions are present in **two alleles** in each individual, regardless of their sex, allowing the application of heterozygosity and homozygosity terminology [18].

Effect of SNPs' Genotypes.

In some cases, having only one chromosome with the minor allele (AB) of a SNP might be detrimental, while in others, both chromosomes with the minor allele (BB) may be needed to observe the detrimental effect [19]. There are scenarios where none of the three genotypes exhibit a harmful effect.

Furthermore, it is important to note that in certain cases, the harmful effect of a genotype at a specific SNP may only manifest when combined with particular genotypes of other SNPs. In fact, recent studies on quantitative traits in humans suggest that genetic variation for many

[3] Hemizygosity also applies when one of the usual two alleles is lost at autosomal loci [17].

traits is primarily influenced by numerous regions in the genome, each contributing a small effect [20]. Understanding these nuances is crucial for unraveling the genetic basis of traits and diseases.

1.4 Data Science Empowering Precision Medicine.

The concept of **precision medicine** combines genomics and healthcare: by guiding healthcare approaches based on an individual's genetic profile, precision medicine aims to optimize treatment effectiveness and minimize adverse effects. The vast volumes of genomic and clinical data available require **sophisticated data analysis** in order to unravel the complex relationships between genes, traits, and diseases. Advanced algorithms spot trends, making connections that can lead to better healthcare outcomes [21].

1.5 GWAS and PGMRA.

A **Genome-Wide Association Study (GWAS)** is a research approach in genomics that involves scanning the entire genome of individuals to identify SNPs associated with particular traits, diseases, or conditions. Since 2005, more than 5,000 GWAS have been published for almost as many traits. These studies have offered insights into the loci and genes underlying phenotypic traits, and are beginning to demonstrate clinical utility by identifying individuals at increased risk for common diseases [22].

However, it has been proposed that SNPs discovered by traditional GWAS analyses account for only a small fraction of the genetic variation of complex traits in human populations [23]. In this context, it is suggested that the majority of the **missing heritability** [4] remains latent within the GWAS data. To uncover such latent information, the **PGMRA server** was developed. PGMRA is capable of finding many-to-many relationships in genomic data and discovering the hidden architecture of a disease. To do so, it can also integrate other domains of knowledge. Subsequently, by incorporating the subject's status post hoc, the risk surface of a disease can be established in an **unbiased manner** [24].

PGMRA significantly overcomes several drawbacks of traditional GWAS analyses, which include challenges such as limited reproducibility, the struggle to identify causal SNPs due to tagged SNPs

[4] Heritability refers to the proportion of variability in a particular trait or characteristic within a population that can be attributed to genetic factors.

not always being causal, difficulties in detecting multiple genetic origins (missing heritability), and an inability to identify epistatic effects [5] [24].

1.6 PGMRA Biclusters.

The algorithm takes a genotype database (SNPs × subjects) as input. The initial phase of the process involves extracting genotype biclusters using **fuzzy biclustering**. The second phase focuses on the **discovery of significant relationships in the biclusters** found in the preceding step, resulting in many-to-many relationships. In the following, we will concentrate on elucidating the first phase, which focuses on the acquisition of high-quality biclusters.

Factorization of the genotype data is performed independently using, by default, a version of a Non-negative Matrix Factorization method (NMF) proposed and termed **Fuzzy NMF (FNMF)**, which allows overlapping among sub-matrices and detection of outliers. This method is applied recurrently to generate multiple clustering results using various initializations with different maximum numbers of clusters and thus avoids any preassumption about the ideal number of clusters. For each run, all clusters are selected composing a family of genotype biclusters $G = \{G1, ..., Go\}$, which may include overlapped, partially redundant, and different sizes of biclusters.

FNMF is a variation of other matrix factorization algorithms which are described below.

1.6.1 Non-Negative Matrix Factorization (NMF).

NMF is a **dimensionality reduction** method in data analysis. It is employed to discover hidden patterns (referred to as "components" or "latent features") within a non-negative data matrix. It can also serve as a **clustering technique**, associating each instance with one of the identified patterns. The fundamental concept is to approximate the original matrix, V, with the product of two lower-rank matrices, W and H, **with the property that all three matrices have non-negative elements**. This non-negativity makes the resulting matrices easier to inspect [25, 26].

Consider a matrix V with c rows and n columns (e.g. c attributes and n instances), denoted as $V_{(c \times n)}$. The outcome of applying NMF, with the number of extracted latent components set to

[5] Epistasis is a phenomenon in genetics in which the effect of a mutation (SNP in this case) is dependent on the presence or absence of other mutations.

k, would result in an approximation of V. This approximation, depicted in Figure 4, is obtained through the multiplication of a matrix $W_{(c \times k)}$ (with as many rows as V and as many columns as the number of latent components being extracted) by $H_{(k \times n)}$ (with as many rows as the number of latent components and as many columns as the number of columns in matrix V) [26].

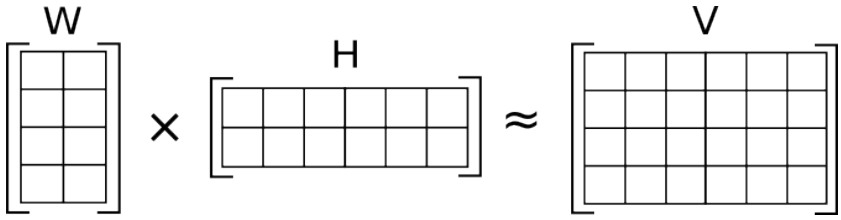


Figure 4. Approximate NMF. The matrix V is represented by the two smaller matrices W and H, which, when multiplied, approximately reconstruct V. In this case, the number of extracted components (k) is 2. Figure taken from a external source [26].

The matrix W is referred to as the **components matrix**, as each of its columns represents a "latent feature" of matrix V (a vector of "prototype" values for each attribute in V). Conversely, the matrix H is termed the **weights matrix** (or transformed data matrix), as each of its columns corresponds to a vector of weights that each of the latent features holds in the instance corresponding to that column. Therefore, the i^{th} column (instance) of V (V_i) is approximated by the product of W and H_i. As a result, the value that the instance has for a single attribute is a linear combination of prototype values (one from each latent feature) in that attribute, weighted by the latent feature weights in that instance [26].

NMF possesses an **inherent clustering property**; in fact, it automatically clusters the columns of the input data. As H stores the weights of each latent feature for every column in V, each instance can be associated with a latent feature, identified as the one with the highest weight for that instance (as simple as determining which row holds the highest value in a specific column of H). Hence, each column of W can be interpreted as a latent feature or as a prototype vector representing the centroid of a cluster of V's instances [26].

There are several ways in which W and H matrices may be found. Lee and Seung's multiplicative update rule has been a popular method due to the simplicity of implementation [26]. It involves iteratively adjusting matrices W and H to approximate the original matrix V as closely as possible, while maintaining non-negativity constraints on all entries [27].

1.6.2 Bio Non-Smooth NMF (bioNMF).

The original bioNMF biclustering method uses the non-smooth variant of the NMF algorithm (nsNMF). This variant achieves an easier interpretation of the components (k) due to the intuitive sparse, non-overlapped part-based representation of the data. To build the biclusters, once the W and H matrices are calculated, the method selects the most representative features and instances for each component. The bioNMF algorithm define as component-specific rows or columns those rows or columns in the W and H matrices, respectively, that show high coefficients for a given component, as well as low coefficients for the remainder components. This is achieved by sorting the rows in W in descending order by their coefficients in a given column, and then, selecting only the first consecutive rows, whose highest value in W is the coefficient in that column. This procedure is repeated for each column of W. Analogously, the process is repeated for the H matrix to select the columns. **The set of selected rows and columns for each component define a bicluster** [24]. It is crucial to emphasize that these biclusters do not overlap with each other.

1.6.3 Fuzzy NMF Biclustering Method (FNMF).

This is the algorithm employed by PGMRA, which is a **fuzzy variation of bioNMF that allows each column or row to belong to multiple biclusters simultaneously**, enabling overlapping. To achieve this, a new parameter called "fuzziness" is introduced to the algorithm. This parameter takes a value between zero and one, representing membership rather than probability. The process now employed to compute the biclusters for each component is as follows:

- The first component of W (first column) is selected, and a threshold is calculated using the following formula:

$$\text{Threshold} = \max(W_{i1}) \times (1 - \text{fuzziness})$$

  All rows of W with values in column 1 greater than or equal to the threshold are selected.
- The first component of H (first row) is selected, and a threshold is calculated using the following formula:

$$\text{Threshold} = \max(H_{1i}) \times (1 - \text{fuzziness})$$

All columns of H with values in row 1 greater than or equal to the threshold are selected.

- The selected rows and columns form the bicluster of the first factor. The same process is repeated with the subsequent components of W.

Fuzziness and outlier detection are connected features as occurs in possibilistic clustering, which avoids the constraint that any observation must belong to at least one bicluster. Low values of fuzziness generates crisp partitions (cohesive biclusters) and emphasizes the detection of outliers; in contrast, high values of fuzziness increments the degree of overlapping, and thus, the algorithm becomes less sensitive to outliers (flexible biclusters). Particularly, when the fuzziness is set to 0, only one row and one column belong to each bicluster, whereas if the fuzziness is set to 1, all the rows and columns belong to every bicluster [24].

1.6.4 Identification of Good Clusters without Determining a priori a Specific Number of Maximum Clusters.

Setting a small number of clusters a priori tends to result in a few large-sized clusters, which can be defined as general clusters that may not fit the data well. Conversely, using a high number of clusters will generate numerous clusters of small size (i.e., finer-grained partitions), which can be defined as specific clusters. These specific clusters might lead to overfitting but they are also likely to reveal true properties of the discovered object [24].

Although there are many validity indices within cluster analysis that suggest the best number of clusters for a given dataset, they often yield contradictory results [24]. **Determining the optimal number of clusters remains an unresolved computational problem, because different characteristics emerge from different assumptions about this number**[28].

Instead of dealing with the diverse and controversial outcomes often obtained from optimizing validation indices that suggest a single number of groups, PGMRA, by default, calculates groups for all partitions generated with maximum number of clusters between 2 and $\sqrt{n}$, where n is the number of observations (subjects). The selection of the best clusters is postponed until all multi-layer partitions are examined, in order to **select a set of clusters that offer an optimal description of the sample**. These clusters can be chosen from different partitions generated by a different number of clusters [24]. Therefore, PGMRA seeks **good clusters rather than a comprehensive grouping** [24] (unlike classic clustering methods, which derive

a global model).

1.7 Colorectal Cancer.

Colorectal cancers (CRCs) encompass two types of **highly aggressive and common types of cancers**, namely, colon and rectal. Globally, while colon cancer is the fourth most common malignancy, rectum cancer is the eighth most common one. Collectively, CRCs present the third most commonly diagnosed form of cancer worldwide, accounting for 11% of all diagnosed cancer cases [29] [6]. Additionally, CRC is the second most lethal cancer worldwide. In this context, the prevalence of CRC cases varies from one geographical location to another, as Hungary has the highest incidence of CRC among males and Norway has the highest incidence among females. Nevertheless, CRC is frequently diagnosed in men from Japan, South Korea, the Middle East region, and Slovakia, with mortality being highest in Saudi Arabia, Oman, and the UAE [29]. Although modern research was able to shed light on the pathogenesis of CRC and provide enhanced screening strategies, **the prevalence of CRC is still on the rise** [29].

1.7.1 CRC Heritability.

Positive family history seems to have a part in approximately 10–20% of all patients with CRC, with varying risk depending on the number and degree of affected relatives and age of diagnosis [31]. Based on twins and family studies, **estimates for heritability of CRC range from 12% to 35%**. Although several GWASs of CRC have successfully identified cancer susceptibility genes that are associated with CRC risk, most factors causing heritability are still elusive and subject to further study [32].

1.7.2 Molecular Pathological Classifications for CRC.

Two molecular pathological classifications for CRC are described [29].

TCGA classification.

The Cancer Genome Atlas (TCGA) project introduced a classification based on integrated molecular analysis which involves two groups:

[6] "Colon cancer" and "colorectal cancer" are often used interchangeably [30]. In this Master's thesis, we will adopt this practice. From now on, whenever we use the term "colon cancer", we are referring to colorectal cancer.

1. **Hypermutated tumors (≈ 16%)**: Tumors accumulating errors in repetitive DNA sequences, called microsatellites, due to malfunctioning repair mechanisms.
2. **Non-hypermutated tumors (≈ 84%)**: These are microsatellite stable (MSS) cancers with:
    - A high frequency of DNA somatic copy number alterations (SCNAs) [7].
    - Dysregulated Wnt pathway [8].
    - And frequent mutations in key genes like: APC, KRAS, PIK3CA, SMAD4 or TP53.

Guinney et al. Classification

On the other hand, Guinney et al. [36] aggregated gene expression datasets using 18 different CRC and described the four consensus molecular subtypes (CMS) of CRC: CMS1 (**MSI-immune**), CMS2 (**canonical**), CMS3 (**metabolic**), and CMS4 (**mesenchymal**). While we will not delve into the biological differences of each subtype, a summary is provided in Figure 5.

[7] SCNAs involve duplications or deletions of specific DNA segments. They are a pervasive trait of human cancers that contributes to tumorigenesis by affecting the dosage of multiple genes at the same time. [33]

[8] The Wnt signaling pathway is a critical mediator of tissue homeostasis and repair, and frequently co-opted during tumor development. Almost all colorectal cancers (CRC) demonstrate hyperactivation of the Wnt pathway, which in many cases is believed to be the initiating and driving event [34]. The mammalian genome includes 19 Wnt genes. [35]

Subtype	Gene Expression	Prognosis
CMS1 (MSI immune)	• Deregulated DNA mismatch repair, MSI and MLH1 silencing • CIMP-high with common *Serine/threonine-protein kinase B-Raf* (*BRAF*) mutations and low number of SCNAs • Immune infiltration and activation	Very poor survival rate after relapse
CMS2 (Canonical)	• Express epithelial signatures with Wingless (Wnt) and *MYC* signaling activation • Frequently exhibit loss of TSGs and overexpression of oncogenes than the other subtypes	Better survival rate after relapse in comparison to other subtypes
CMS3 (Metabolic)	• Fewer SCNAs as compared to CMS2 and CMS4 • Epithelial signatures • Metabolic dysregulation in a variety of pathways with frequent *KRAS* mutations • Slightly higher presence of CIMP-low	Better survival rate after relapse in comparison to other subtypes
CMS4 (Mesenchymal)	• Activation of Transforming growth factor-β (TGF-β) • Upregulated expression of EMT genes • Enhanced expression of genes regulating inflammation, matrix remodeling, stromal invasion and angiogenesis	Worse overall and relapse-free survival as compared to other subtypes

CIMP: CpG island methylator phenotype; EMT: epithelial-mesenchymal transition; MSI: microsatellite instability; SCNA: somatic copy number alterations.

Figure 5. Biological differences in the gene expression-based molecular subtypes of CRC. Figure taken from an external source [29].

1.7.3 Risk Factors.

CRC pathogenesis is highly complex and diverse, and is induced by several risk factors including sporadic, familial, and inherited [29].

- **Sporadic cases comprise 70%** of CRC cases and are caused by environmental and dietary factors (cigarette smoking, excessive alcohol consumption, sedentary lifestyle, obesity, and diets high in fat and low in fiber).
- **Familial CRC cases comprise 25%** of the cases and affect individuals with family history of CRC.
- **Genetic or inherited cases account for 5–10%** of the cases and are categorized based on the presence or absence of colonic polyps.

Other risk factors include **being male** (with incidence and mortality rates significantly higher in males than in females [37, 38]) or the presence of **existing diseases** such as **long-standing inflammatory bowel diseases (IBD)** [29].

1.7.4 Metastasis in CRC.

Tumor metastasis is the movement of tumor cells from a primary site to progressively **colonize distant organs** [39]. Most cancer-associated deaths occur due to metastasis, yet our understanding of metastasis as an evolving, heterogeneous, systemic disease and of how to effectively treat it is still emerging. Metastasis requires the acquisition of a succession of cellular traits to disseminate, variably enter and exit dormancy, and colonize distant organs. The success of these events is driven by clonal selection [9], the potential of metastatic cells to dynamically transition into distinct states, and their ability to co-opt the immune environment [40].

The metastatic dissemination of primary tumors accounts for about 90% of all colon cancer deaths. However, there are almost no prevalent mutations conclusively associated with metastatic colon cancers. Instead, specific characteristics of the tumor microenvironment, such as diminished immune cytotoxicity, are indicative of unfavorable outcomes in CRC patients [41].

1.8 Project's Motivation and Objectives.

Despite being diagnosed with the same type of cancer, CRC patients may share few common symptoms, exhibit varying degrees of severity, and respond differently to treatments [42]. The heterogeneity of this disease is likely conditioned by the genetic variants of each patient. Thus, the main objective of this master's thesis is to evaluate the ability of the **PGMRA algorithm** to generate new insights in the **genetic characterization of CRC patients**, particularly within a **limited sample size**.

The analysis of genotypic SNP data using PGMRA will enable the identification of **fuzzy biclusters (CRC patients $\times$ SNPs) in an unsupervised manner.** The whole set of patients or SNPs included in these biclusters **can be studied collectively**, allowing for the examination of both clinical and genetic characteristics of the complete sample. However, the analysis of patients and SNPs for **each bicluster separately** will enable the identification of genetic differences among subgroups of colon cancer patients that may be associated with clinical or molecular disparities of the disease. Both types of approaches (bicluster-no-specific and bicluster-specific) will be carried out to contrast the obtained results.

[9] Clonal selection is the process of proliferation of descendant cells with specific genetic or functional characteristics from a single stem cell.

The expectation is that among these features that seem to be linked to the disease, some are already documented in the literature (and thus serve as **external validation** of the biclusters' identification and the pipeline followed for its analysis), while others are not covered in the literature and may represent the **extraction of novel knowledge.** The **specific objectives** of the project are described below.

Objective 1. Identification of good genotypic biclusters in an unsupervised way using PGMRA.

PGMRA will be used to analyze a SNP genotyping cancer study, which includes patients with different types of cancer diagnoses (colon, melanoma, breast, and control), aiming to generate genotypic fuzzy biclusters (patients $\times$ SNPs) in an unbiased manner (as previously explained, the identification of these biclusters will not consider the subject's clinical status (case/control)). Subsequently, those **biclusters containing exclusively colon cancer patients will be selected.** These biclusters will identify genetic variants relevant to different subject subgroups, thus suggesting possible distinct pathways for disease origin or disease syndromes.

Objective 2: Analysis of biclusters' similarity regarding patient and SNP composition.

The goal is to comprehend how patients and SNPs from the entire dataset are distributed among the biclusters created by PGMRA. Specifically, the study will focus on how these biclusters differ in the number of patients, the number of SNPs, their hierarchical relationships, etc. Special emphasis will be placed on determining the percentage of shared patients and SNPs for each bicluster pair.

Objective 3. Analysis of the complete sets of CRC patients and SNPs.

A comprehensive data analysis encompassing patients and SNPs from all CRC biclusters will be conducted. This analysis includes:

- Study of **clinical attributes** among patients (through graphical representations) and exploration of clinical values **tendency to co-occur** (using association rules).
- Analysis of **key attributes in SNP annotation.**
- Identification of **intergenic SNPs** and their **potential regulatory roles**, particularly in the colon.

- Identification of **genic SNPs** and analysis of their impact.
- Analysis of **affected genes' functions** and their potential association with the disease. Particularly, the following operations will be carried out:
  - **Functional gene annotation** to uncover direct or indirect associations with colon cancer or cancer in general.
  - **Over-representation analysis** exploring affected pathways and diseases.
  - Verification of possible **differential gene expression in CRC** compared to healthy colorectal tissue, using external RNA-seq data.
  - Creation of a CRC representative **protein-protein interaction network** from the obtained knowledge and identification of central nodes and communities, using network analysis algorithms.

Objective 4. Bicluster-specific analysis.

An analysis similar to the overall one will be conducted for each of the biclusters, pursuing the following targets:

- **Identification of patients and SNPs** from the specific bicluster.
- **Replication of most of the previously mentioned operations**, including the analysis of clinical and SNP attributes, the exploration of the regulatory function of intergenic SNPs, the impact of genic SNPs, gene set over-representation analysis, etc.
- **New analytical operations**:
  - Creation of a **genetic map for each gene** affected in the bicluster, displaying the bicluster-specific SNPs. This information will be analyzed in conjunction with the gene annotation-data extracted in the global analysis.
  - Calculation of **SNP genotypic frequencies** within the bicluster, along with the **frequency of the minor allele**.

An additional goal consists on the generation of visualizations such as **heatmaps** to enhance the comprehension of **genotypic differences among patients from distinct biclusters**.

CHAPTER 2: INPUT DATA AND METHODOLOGY

2.1 Data Description and Methodology Flowchart.

Generation process of the input data for this study.

1. The study cohort consisted of 266 individuals categorized into control, colon, melanoma, and breast cancer groups.
2. **Genotyping data** for each patient were obtained using **Illumina microarray technology.** This technology is summarized in Appendix B.1.
3. The generated data was processed using **GenomeStudio Illumina** [10] and subsequently converted into **PLINK files** [11].
4. Out of the initial 266 subjects, **264** successfully passed the quality control filter. Subsequently, SNP filtering was performed at a significance threshold of 0.001, resulting in a final dataset containing **17,314** SNPs out of the original 654,027. In this dataset, genotypes were numerically encoded as follows:
   - **Homozygous Major** (AA) ≡ *1*
   - **Heterozygous** (AB) ≡ *2*
   - **Homozygous** Minor (BB) ≡ *3*
   - **Missing value** ≡ *0*
   - Initially, we were unaware of the **hemizygosity** encoding.
5. **PGMRA** was executed on these PLINK files. Specifically, PGMRA was performed on all patients with available genotype data (n=264) and all variables (n=17,314), regardless of their classification as control, colon, melanoma, or breast cancer. Executions were performed with a variable number of biclusters, ranging from 2 to 17.
6. Out of the 152 total biclusters obtained, those exclusively containing CRC samples were selected. A total of **42 biclusters** were identified. Remarkably, these groups together included **77 colon patients** out of the total with available genotype data (n=78) and **3,915 SNPs**. This approach enabled the **identification of distinguished features between colon cancer and other cancer types as well as controls.**
7. SNPs were searched for in **Ensembl** [43]. Ensembl is a genomics database and analysis

[10] GenomeStudio Illumina is a software platform developed by Illumina for the analysis and interpretation of genotyping data obtained from Illumina microarray technology.

[11] PLINK is a widely used software tool for the analysis of genetic data.

tool, offering valuable information on genes, transcripts, and regulatory elements across various species.

As a result of this process, the following datasets were obtained:

- **Genotypic dataset.** This table originally encompassed the data on which PGMRA was executed to extract biclusters. However, it was filtered with the 3,915 SNPs (rows) and 77 CRC patients (columns) belonging to the selected CRC-biclusters. The values indicate the encoded genotype (*1, 2, 3, 0*).
- **Membership matrix of SNPs in biclusters.** Rows represent all SNPs. Columns represent all biclusters, not just the selected (CRC-specific) ones. A value of "1" indicates the presence of the SNP in the group, and "0" indicates absence.
- **Membership matrix of patients in biclusters.** Rows represent all subjects. Columns represent all biclusters, not just the selected ones. A value of "1" indicates the presence of the subject in the group, and "0" indicates absence.
- **Patient Clinical dataset.** Clinical data of subjects included in the selected groups. Additionally, this table incorporates the data from the membership matrix of patients in biclusters.
- **SNP Ensembl dataset.** Information on SNPs extracted from Ensembl. Additionally, this table incorporates the data from the membership matrix of SNPs in biclusters.

Basically, as shown in Figure 6, the main table is the genotypic table, containing encoded genotypes for each SNP and patient. To explore additional characteristics of these SNPs and patients, we will use the SNP Ensembl dataset and the patient clinical dataset. Additionally, these tables incorporate the data from the membership matrices of SNPs and patients, respectively, in biclusters. This information allows for the indexing of both the genotypic table and the clinical or Ensembl table using bicluster-specific patient and SNP identifiers. As mentioned when describing the objectives, both the analysis of the entire datasets and the analysis of the datasets indexed for each bicluster are two approaches pursued in this project.

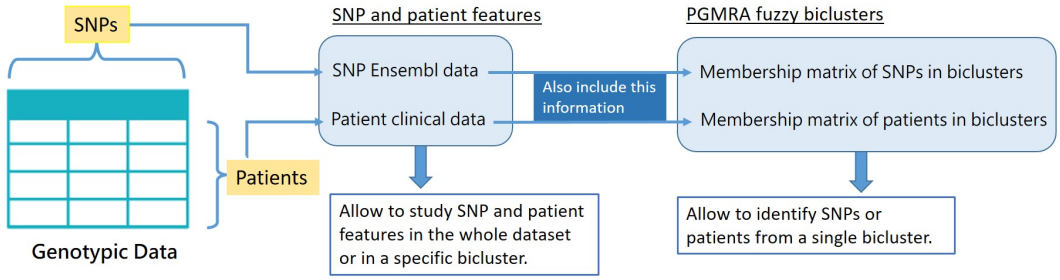


Figure 6. Data for analysis.

Methodology flowchart.

The methodology followed for the analysis of the datasets presented in Figure 6 is summarized in Figure 7. The following sections of this chapter are dedicated to separately describing each of the methods carried out.

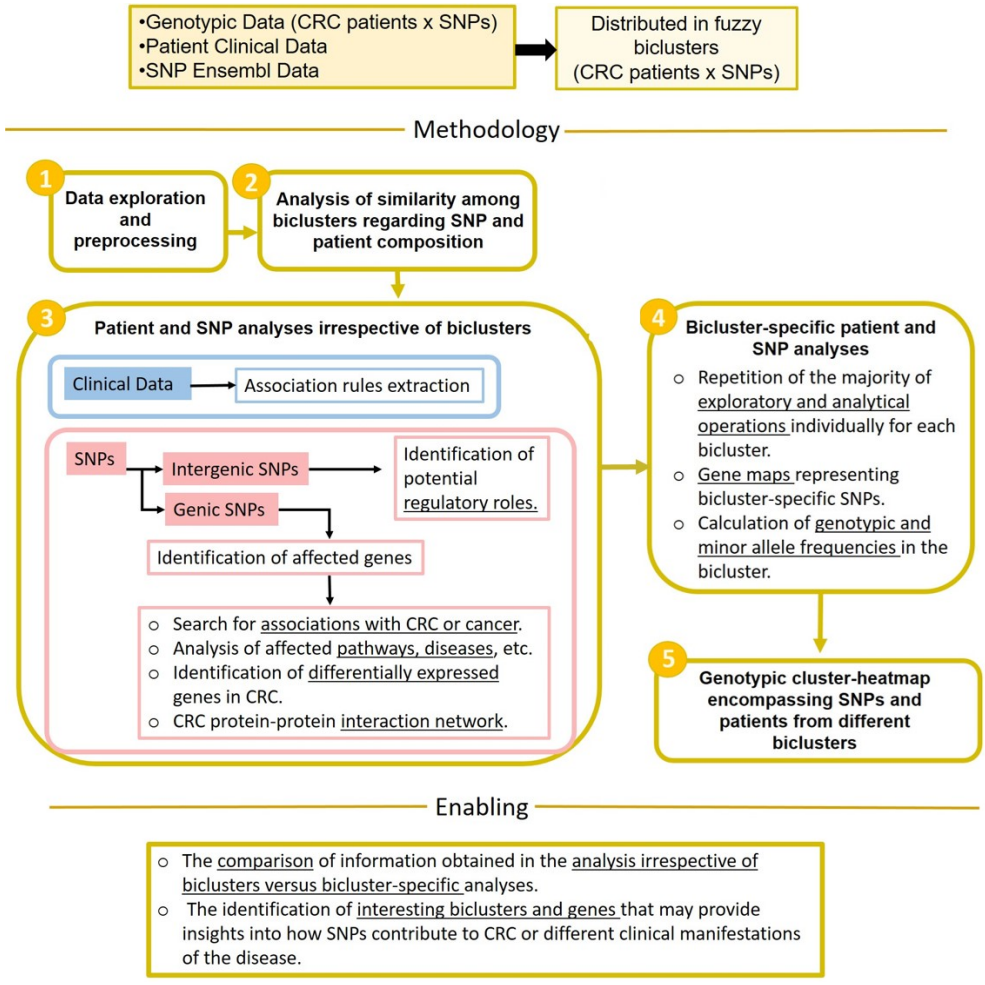


Figure 7. Methodology.

2.2 Data Exploration and Preprocessing.

2.2.1 Detection of Erroneous Patient Samples and Subsequent Filtering of Patients, Biclusters, and SNPs.

Upon creating genotypic cluster-heatmaps for initial testing, it became apparent at first glance that there were **two pairs of patients sharing the same or very similar genetic patterns**. Further examination confirmed that within each pair, both patients shared exactly the same genotype for all SNPs. Moreover, the genotypic pattern of these four patients appeared quite **atypical** compared to the rest. In fact, 26 out of the 42 CRC-specific biclusters exclusively contained either one of these pairs of patients or all four together. We concluded that these samples were erroneous and therefore decided to **exclude the four patients** from the study.

The removal of the four samples, along with the biclusters exclusively associated with them, and the SNPs belonging solely to those biclusters was carried out in the genotypic, clinical, and Ensembl tables. This constituted the first preprocessing step for these datasets. As a result, the new genotypic table contained **73 CRC patients and 2,496 SNPs, relative to 16 biclusters.** Therefore, the focus of this research is on the **study of these 16 biclusters**.

2.2.2 Missing Values.

After the elimination of the four erroneous samples, we examined the missing values in the genotypic table in Python [12], as well as in the clinical table. For each table, the percentage of missing values was calculated in both columns and rows. The results were sorted.

We did not perform this analysis on the Ensembl table because the missing values in this dataset result from SNPs not recognized by Ensembl (with missing values in all columns) and intergenic SNPs (with missing values in attributes related to genes).

[12] The missing values in the genotypic table are encoded with the number *0*, so for analysis, we changed them to the value *numpy.nan*

2.2.3 Data Engineering on the Clinical Dataset and Visualization of Attributes Distribution.

After generating bar charts to visualize the distribution of clinical attributes across the 73 patients, we opted to conduct various data engineering operations on the clinical dataset, including column removal, grouping of categories, discretization, etc. (a more comprehensive description of the steps followed can be found in Appendix B.2). Subsequently, we repeated the creation of bar charts.

2.2.4 Preprocessing of the Ensembl Dataset and Visualization of Attributes Distribution: Distinguishing Genic and Intergenic SNPs.

The Ensembl dataset contains annotations for each of the SNPs used in this project. These annotations were retrieved using the reference genome GrCh37 (Genome Reference Consortium human genome assembly 37). Each row corresponds to a SNP, and each column represents an attribute (chromosome, possible alleles, minor allele, etc.). As previously mentioned, additional columns are included at the end, corresponding to the PGMRA biclusters and indicating, with boolean values, the presence or absence of each SNP in each bicluster.

We observed that in the **"chromosome" column**, not only autosomal chromosomes (1-22) and sex chromosomes (X and Y) were present but also many other names, describing **patches** (specific segments of the reference genome that have been updated compared to the previous version of the genome) [13]. What we did was to keep only the rows annotated with a standard chromosome name (a number between 1 and 22, or "X" or "Y"). In this operation, **SNPs not recognized by Ensembl were also removed**, characterized by having missing values in all Ensembl annotation fields.

The resulting table included genic SNPs (those with a value in the "ensembl_gene_stable_id" column) and intergenic SNPs (those with an empty string in this column). We created a **subset of the table for each type of SNP (genic and intergenic)**. The reason is that both types require different preprocessing steps to obtain unique rows of SNPs, as we will see below.

Preprocessing of Genic SNP Ensembl Data.

[13] All of the SNPs located in patches also appeared, in another row, annotated with the standard chromosome name.

Many genic SNP can have different consequences at the transcript level, as detailed in the "consequence_type_tv" attribute of the Ensembl table. To simplify analysis, **only the most severe consequence for each SNP was kept**. To achieve this, we extracted from Ensembl a list of the possible consequences, ordered from highest to lowest severity [44]. This preprocessing was made with tailor-made R scripts.

Additionally, this Ensembl list also provided the correspondence between each type of SNP consequence and the **prediction of its impact** ("HIGH", "MODERATE", "LOW" or "MODIFIER"). We added this new information to the genic SNP Ensembl dataset. Finally, **bar charts** were generated to observe the distribution of the most informative attributes.

Preprocessing of Intergenic SNP Ensembl Data.

In the intergenic SNP Ensembl table, most rows have the **"associated_variant_risk_allele"** column empty. However, some SNP identifiers appear **multiple times** in different rows due to having different values in this column. Since this column was not going to be used, the decision was made to take the **first row for each repeated SNP**. In this case, we also created a **bar chart**, specifically for the "chromosome" attribute, as columns related to the affected gene are useless (and empty) for intergenic SNPs.

2.2.5 Exploring Hemizygosity Encoding.

Considering that the values of the genotypic dataset were supposed to describe homozygosity and heterozygosity, we raised questions about **how genotypes of SNPs falling on the sex chromosomes were encoded.** On one hand, men have only one X chromosome, so they cannot be either homozygous or heterozygous for X chromosome SNPs (they are necessarily hemizygous). On the other hand, women do not have a Y chromosome and thus cannot have genotypes associated with SNPs from this chromosome.

To address this, we generated subsets of the genotypic table that could clarify our queries:

- Subdataset of male patients and X chromosome SNPs.
- Subdataset of male patients and Y chromosome SNPs.
- Subdataset of female patients and Y chromosome SNPs.

2.3 Analysis of Similarity Among Biclusters Regarding SNP and Patient Composition.

Since the PGMRA biclusters are fuzzy (overlapping), the same SNP or patient can belong to multiple (even all) biclusters. To assess the extent of overlap between each pair of biclusters regarding their composition of SNPs and patients, we generated percentage coincidence matrices and graphs based on these matrices. We also observed the differences in the number of patients and SNPs among biclusters.

2.3.1 Analysis of Patient Sharing Among PGMRA Biclusters.

Patient coincidence matrix.

We computed the percentage of shared patients between each bicluster and all others. These results were then used to fill a 16×16 matrix in which each entry signifies the proportion of patients from one bicluster (row) who are also part of another bicluster (column). To provide additional information, the **diagonal elements were filled with the number of patients** from the bicluster that appears both in the corresponding row and column. For improved visualization, a color scale was used to represent the magnitude of the percentage in each matrix cell.

Graph Generation from the Coincidence Matrix.

We saved the coincidence matrix without the diagonal and exported it to **Gephi**, a network visualization and analysis software [45], as an adjacency matrix, to create a **weighted directed graph**. In this graph, the nodes represent the 16 biclusters, and the weight of an edge between two nodes corresponds to the percentage of patients from the source node (bicluster) that are also included in the target node. The purpose of constructing this network is not to precisely display the coincidence percentages (as that is accomplished by the matrix), but to visualize and reflect the **nodes with higher centrality** through the graph's topology.

To achieve this, we utilized various Gephi options, ultimately arriving at the following configuration:

- **Nodes.**
  - Color: determined by eigenvector centrality (darker shades indicate higher centrality).

  - Size: based on the number of patients (larger sizes represent more patients).
- **Edges.**
  - Color: arbitrary.
  - Size (including label size): determined by edge weight.
  - Label: edge weight. Displayed only for weights greater than 65.
- **Layout: circular**, with nodes arranged in a clockwise direction based on decreasing betweenness centrality values.

Explanations for these terms related to graph theory (node centrality, eigenvector centrality and betweenness centrality) can be found in Appendix B.3, which may provide a better understanding of the graph's purpose.

2.3.2 Analysis of SNP Sharing Among PGMRA Biclusters.

The procedure is identical to that described for studying patient sharing.

2.4 Extraction and Filtering of Association Rules in the Clinical Dataset.

Association rules have been one of the most widely used data mining techniques to extract valuable knowledge from large databases. They are employed to identify and represent **dependencies between elements (items)** in a database where the class to which the data belongs is unknown (unsupervised learning). Association rules are rules that have the following format: $X \Rightarrow Y$, where X and Y are sets of items satisfying $X \cap Y = \emptyset$ [46].

In the context of this master's thesis, we aim to extract association rules from the clinical dataset. Here, the **items** to be associated are the **pairs "attribute: value"** from the table, and the **transactions**, which represent cases of joint occurrence of items, correspond to the **patients**, meaning each row of the table.

Classic measures of association rules.

Support. The support of an item is the frequency of the item appearing in the database, and the support of an association rule $X \Rightarrow Y$ is the frequency of the itemset $X \cup Y$ in the set of transactions. This measure takes values in the range [0.0, 1.0]. A support of 1.0 indicates that it appears in all transactions in the database, and 0 means it doesn't appear in any. An itemset is termed "frequent" when it surpasses the minimum support threshold chosen by the expert.

Confidence is a measure of fulfillment for an association rule. It is defined as follows:

$$confidence(X \Rightarrow Y) = support(X \Rightarrow Y)/support(X) \tag{1}$$

This measure also takes values in the range [0.0, 1.0]. A confidence of 1.0 means that whenever X occurs, Y also occurs, and 0 indicates that when X occurs, Y does not.

Extraction of association rules.

The extraction of association rules from a database involves generating rules that meet the minimum support and confidence thresholds. We used a classic method, the **Apriori algorithm**, that addresses this problem by reducing the set of candidate itemsets (frequent itemsets candidates to form a rule) [47]. A brief description of the algorithm's foundation can be found in Appendix B.4.

Filtering rules using a metric that penalizes high support in the consequent.

Among the generated rules, it is common to find many with an item in the **consequent that frequently appears** in the database, indicating a high support. This situation **diminishes the utility of the rule**, as not only does the antecedent of the rule frequently co-occur with the consequent, but any item in the database does as well. Since the confidence metric doesn't consider the support of the consequent, we employed another metric that penalizes such cases: **lift**.

$$Lift(X \Rightarrow Y) = \frac{Conf(X \to Y)}{Sop(Y)} = \frac{Sop(X \to Y)}{Sop(X) \cdot Sop(Y)} \tag{2}$$

When the lift is greater than 1, it indicates that items X and Y are more likely to appear together than apart, which means that the presence of one item increases the likelihood of the other, signifying a positive correlation (not causation) between them. A value of 1 indicates statistical independence, meaning that the antecedent and consequent are not correlated, while values lower than 1 indicate a negative correlation [48].

Parameters configuration.

To generate association rules from the clinical dataset, we employed the **R package arules** to

execute the Apriori algorithm with a **minimum support threshold of 0.05** and a **minimum confidence threshold of 0.7**. This allows to detect rules with relatively high confidence even when their itemsets do not have a high frequency of occurrence (hence the low support threshold). Only rules with **lengths ranging from 2 to 4 items** were generated. Subsequently, **rules with a lift smaller than 1.4 were filtered out**. Finally, we **manually selected** some of the rules that appeared to be more interesting.

2.5 Identification of Interesting Intergenic SNPs Using RegulomeDB.

Nearly 90% of the disease risk-associated SNPs identified by GWASs are in non coding regions of the genome. The annotations obtained by analyzing functional genomics assays can provide additional information to pinpoint causal variants, which are often not the lead variants identified from association studies. However, the **lack of available annotation tools** limits the use of such data [49]. To address the challenge, the **RegulomeDB database** and **RegulomeDB web server** (`http://regulomedb.org`) were built to prioritize and annotate variants in non-coding regions.

RegulomeDB annotates a SNP by intersecting its position with genomic intervals identified from functional genomic assays and computational approaches. It also incorporates variant hits into a heuristic ranking score, representing its potential to be functional in regulatory elements.

The RegulomeDB web server **prioritizes the query variants by functional prediction scores** shown in a sortable table. For any variant of interest, an information page on **five types of supported genomic evidence**, as well as a **genome browser** view is displayed. Each of the six sections can be clicked to show more detail for functionality exploration.

In the data analyzed in this Master's Thesis, there are **1,025 intergenic SNPs**. Due to the RegulomeDB web server's limitation on query length, we had to input the identifiers of these SNPs in **five separate searches**. We unified the results from the five resulting tables and conducted the analysis in R. As a result of this analysis, we obtained a dataset containing the following attributes: **"regulatory_prob"** (the probability that the SNP has regulatory function in general), **"ranking"** (a score given to each SNP based on its supporting evidence [14]), **"chrom"** (chromosome where the SNP is located) and three tissue columns **("colon",**

[14] In general, if more supporting data is available, the higher is its likelihood of being functional and hence the SNP receives a higher score (with 1 being higher and 7 being lower score). The detailed meaning of the ranking

"intestine", and "large intestine") representing the SNP's probability of having a **tissue-specific** regulatory function.

After several filtering operations and observing that no SNP had a probability greater than 0.3 in any of these three tissue-specific fields, we decided to **retain rows with a probability greater than 0.25 in at least one of these three fields**. Later, we also studied which intergenic SNPs from each bicluster belonged to this set of filtered SNPs.

2.6 Gene Annotation.

VarElect.

VarElect (`https://varelect.genecards.org/`) is an advanced tool for **prioritizing genes associated with diseases and phenotypes**. It leverages comprehensive biological knowledge databases to provide context and supporting evidence, aiding researchers in making informed decisions in deep sequencing analyses. VarElect is supported by the GeneCards Suite Integrated Biomedical Knowledgebase, which includes GeneCards, MalaCards, and LifeMap Discovery.

GeneCards.

GeneCards (`https://www.genecards.org/`) is a searchable, integrative database that provides comprehensive, user-friendly **information on all annotated and predicted human genes**. The knowledgebase automatically integrates gene-centric data from $\tilde{1}50$ web sources, including genomic, transcriptomic, proteomic, genetic, clinical and functional information.

2.6.1 Search for Associations with Colon Cancer.

In the data we analyzed in this project there are 972 genic SNPs. Some of them are found within the same gene, resulting in **695 different affected genes** in the set of 73 patients. The 695 Ensembl identifiers (obtained from the genic SNP Ensembl table) were used as **queries in VarElect** to obtain associations of these genes with the **target phenotype**, **"colon cancer"**. We downloaded the XLSX file containing the results.

This XLSX file includes two tabs: one with **genes directly related** (according to VarElect) to the target phenotype and another with **genes indirectly related** to the phenotype (via other

scores can be seen in Appendix B.5, Figure 42.

gene(s) directly related to the phenotype). VarElect's website also provided a list of **100 genes for which it did not find any association** (neither direct nor indirect) with the phenotype. We saved this list. In addition, there were **17 Ensembl identifiers that VarElect could not recognize** (nor translate into known aliases). The search for information about these genes was conducted towards the end of the annotation process.

Genes Directly Related to CRC.

From the table of genes directly related to CRC, we retained only three columns (**"Symbol", "Description", "Category"**) and added four new columns:

- **"General Info"**: to store basic information about the gene's function.
- **"Colon Related Info"**: to capture any evidence of association with colon cancer as well as the colon itself.
- **"Cancer Related Info"**: to record any evidence of gene-cancer associations.
- **"Cancer Census":** a binary indicator, "Yes" or "No", to denote whether the gene is included in the cancer census.

One of the most tedious tasks of this master's thesis involved filling each field for every gene in this table. In most cases, each cell was filled by quoting sentences from articles or databases, in a search guided by the output displayed on the **VarElect website** or GeneCards. A more detailed explanation of the steps followed can be found in Appendix B.6.

Something important to keep in mind is that VarElect serves as a guide for annotating gene-phenotype associations, but it can sometimes be inaccurate. For instance, some genes are directly linked to the disease through evidence in scientific articles, yet VarElect may not recognize these associations. Conversely, there are instances where VarElect identifies a gene as being related to the phenotype based on text mining of articles, even though the association may be nonexistent (for example, if a gene shares an acronym with a drug used for colon cancer, VarElect may mistakenly associate the gene with the disease). As a result, it was **not always possible to fill the "Colon Related Info" field**. In such cases, the gene was removed from this table and its identifier was added to the list of genes not associated with colon cancer for subsequent searches.

Genes Indirectly Related to CRC.

In this table, there is one row for each gene pair: **implicated gene** (indirectly related to the disease) - **implicating gene** (associated with the implicated gene and directly related to the disease). Thus, if a gene is indirectly related to the disease through five genes, that gene appears in five rows, each with a different implicating gene. The data in this table underwent manipulation in both R (for translating **external names to Entrez identifiers** using the org.Hs.eg.db package) and Python (to obtain **NCBI gene summaries** using the Bio Entrez package). Finally, the table was transformed so that instead of having one row for each related gene pair, there is **one row for each implicated gene**. The final fields are: "Implicated Gene", "Implicated Gene Summary", "Implicating Genes" (all genes that indirectly link the implicated gene to the disease are listed in the same cell), and "Implicating Genes Summaries" (all NCBI summaries of the implicating genes in a row appear in the same cell).

2.6.2 Search for Associations with Cancer in General.

For genes that had no direct or indirect associations with colon cancer, a new search in **VarElect** was conducted. This time, the search focused on associations with a **broader phenotype** of interest: **"cancer"**. The process to obtain tables of genes directly and indirectly related to cancer was the same as described for colon cancer.

2.6.3 Manual Search for Associations with Colon Cancer or Cancer in General for Genes that VarElect did not Link to Either of the two Phenotypes.

Finally, both the genes whose **identifiers were not initially recognized** by VarElect and those for which **no associations were found** with either colon cancer or cancer in general were individually examined (in GeneCards and relevant articles). The goal in these cases was to manually search for and annotate potential relationships (with either of the two phenotypes of interest) that VarElect may not have considered. Genes for which an association was found were included in the corresponding table, while those without any identified associations were included in a **fifth table**, containing only the gene name and NCBI summary.

2.7 Gene Set Over-Representation Analysis.

DisGeNET [50] is an integrative and comprehensive resource of gene-disease associations from several public data sources and the literature [51].

Gene Ontology [52] defines concepts/classes used to describe gene function, and relationships between these concepts. It classifies functions along three aspects [51]:

- **MF**: Molecular Function (molecular activities of gene products).
- **CC**: Cellular Component (where gene products are active).
- **BP**: Biological Process (pathways and larger processes made up of the activities of multiple gene products).

KEGG [53] is a collection of manually drawn pathway maps representing molecular interaction and reaction networks. These pathways cover a wide range of biochemical processes that can be divided into 7 broad categories (Metabolism, Genetic information processing, Environmental information processing, Cellular processes, Organismal systems, Human diseases and Drug development) [51].

Over-Representation Analysis.

Over-Representation Analysis (ORA) [54] is a widely used approach to determine whether known biological functions or processes are over-represented (enriched) in an experimentally-derived gene list.

The **p-value** can be calculated by **hypergeometric distribution**:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}} \tag{3}$$

In this equation, N is the total number of genes in the background distribution, M is the number of genes within that distribution that are annotated to the concept of interest, n is the size of the query gene set and k is the number of genes within that set which are annotated to the concept [55]. The background distribution by default is all the genes that have annotation. P-values should be adjusted for multiple comparison [51].

To perform **over-representation tests for DGN diseases, GO terms, and KEGG pathways** on the entire set of affected genes, as well as later on genes directly related to CRC or

bicluster-specific genes, we utilized the *enrichDGN* function from the DOSE package, and the *enrichGO* and *enrichKEGG* functions from ClusterProfiler. These **parameters** were used for the configuration of all functions:

- *pAdjustMethod* = "bonferroni"
- *pvalueCutoff* = 0.05
- *qvalueCutoff* [15] = 0.1

2.8 Gene Expression Analysis Using External RNA-Seq Data.

In order to determine wether any of the genes affected by SNPs in our dataset exhibit differential expression in colorectal cancerous tissue compared to colorectal healthy tissue, we conducted a RNA-seq [16] analysis using real samples from TCGA database related to CRC patients.

TCGA Data Downloading.

TCGA (`https://www.cancer.gov/ccg/research/genome-sequencing/tcga`) is a publicly available database that contains a large collection of genomic and clinical data related to various types of cancer. We accessed TCGA data through the GDC (Genomic Data Commons) portal (`https://portal.gdc.cancer.gov/`). Specifically, **colon, rectum and rectosigmoid junction** were selected as primary sites, as part of different TCGA-projects. Only **"primary tumor" and "solid tissue normal"** sample types were selected [17]. From that, all **STAR-Counts files** included as **Gene Expression Quantification** data types were downloaded.

TCGA Data Preprocessing.

KnowSeq is a freely available R/Bioconductor package for **RNA-seq data analysis**. It is designed to provide comprehensive results for transcriptome analysis, including gene expression quantification, differential expression analysis, and functional annotation [57]. In this project, it was used to conduct preprocessing and subsequent analysis of the downloaled TCGA files.

The preprocessing of TCGA data began with **undersampling of the cancerous class** to diminish the unbalance of the original dataset (see Table 1). Then, using KnowSeq functions, the

[15] The q-value is an alternative approach for calculating False Discovery Rate (FDR) control in multiple testing, similar to an FDR-adjusted p-value.

[16] RNA sequencing (RNA-seq) is a genomic approach for the detection and quantitative analysis of messenger RNA molecules in a biological sample and is useful for studying cellular responses [56].

[17] Samples of healthy colorectal tissue were obtained from the same patients from whom samples of cancerous tissue have been collected. However, for many patients, only the cancerous tissue sample was taken

following steps were performed: count data was transformed into **gene expression values** for each sample; **outliers** (samples of questionable quality due to having an expression distribution significantly different from the rest) were detected and removed and, finally, surrogate variable analysis (**SVA**) model was applied to address **batch effects**.

Table 1 shows the number of samples of each class downloaded from GDC, as well as the number remaining after undersampling and after the elimination of outliers.

Table 1. TCGA Data: Downloaded, randomly selected (undersampling) and filtered samples of each class.

Class	Description	Downloaded	Rand. Selected	Quality samples
HEALTHY	Healthy colorectal tissue	54	54	53
TUMOR	Colorectal cancer	790	200	191

DEGs identification.

The whole sample set was subjected to KnowSeq *DEGsExtraction* function, which performs an analysis to extract DEGs in the two classes of interest. This function was configured with a **p-value** of 0.001 and a **minimum log fold change** (LFC) of 1.5 [18].

Then, we determined **which of our 695 genes belonged to this set** of extracted DEGs (also determining whether they were upregulated or downregulated). To achieve this, we paid particular attention to gene aliases: Appendix B.7 outlines the methodology employed to handle three types of gene aliases, with the aim of identifying any correspondence between our gene names and those of the extracted DEGs. Finally, with the DEGs that were present in our dataset, we created **boxplots and heatmaps** to display their expression across each sample class.

2.9 Minimum Protein-Protein Interaction Network Derived from All Affected Genes.

Proteins rarely operate in isolation; instead, they interact with each other to carry out specific functions [58]. In order to gain an overview of the functionality of proteins encoded by

[18] Log fold change (LFC) is a measure that quantifies the relative difference in gene expression between two conditions or groups. A LFC of 1.5 indicates that the gene has approximately $2^{1.5}$ times higher expression in one condition compared to the other.

the 695 genes affected in the 73 patients and to understand their relationships, we sought to explore the protein-protein interactions among them using a **graph-based approach**.

To achieve this, we inputted the genes' Ensembl identifiers into **NetworkAnalyst** (`https://www.networkanalyst.ca/`), a web service that offers integrative tools for the analysis and visual exploration of protein networks. Specifically, we constructed the **minimum graph of general protein-protein interactions, with a one-degree separation**, using the IMEx Interactome database. In this network, not all the 695 introduced genes appeared, while new nodes were included, representing genes that interact (at the protein level) with more than one seed gene that we provided.

To analyze the graph, we performed the following steps:

- We conducted **node set enrichment** operations using NetworkAnalyst tools, incorporating GO terms, KEGG pathways, etc.
- We exported the network in JSON format, to open it with **Cytoscape**, an open-source software environment for the large scale integration of molecular interaction network data [59].
- We also exported the **seed node names** (which are external names) from NetworkAnalyst to work with them in R, where we performed the following operations:
  - We translated these names back into **Ensembl format** using the *GetGenesAnnotation* function from the KnowSeq package. [19]
  - For each of this Ensembl identifiers, we computed two attributes related to the PGMRA data: the **number of SNPs** falling within the gene and the **number of biclusters** containing at least one SNP in that gene.
  - Using these values, we generated an **attribute table specific to the seed nodes** and imported it into Cytoscape to merge with the node attribute table created by NetworkAnalyst.

Next, we manipulated the network in both **Gephi** (after exporting the graph in GraphML format from Cytoscape) and **Cytoscape**. Gephi was used for enhancing visualization, applying the desired layout, and employing the **Louvain community detection algorithm** (with a resolution of 1.7). In Cytoscape, we conducted **enrichment** analyses for each of the Louvain

[19] Some genes had to be manually translated, using GeneCards data, because KnowSeq did not recognize their external name.

communities obtained with Gephi. We created a table listing over-represented terms manually selected for each community.

Below, the configuration of the graph visualization in Gephi is described:

- **Nodes.**
  - Color: blue (fixed tone) for non-seed nodes and a scale of reds for seed nodes, where darker shades indicate a higher number of biclusters.
  - Size: based on eigenvector centrality (larger sizes indicate higher eigenvector centrality). It also applies to the node label size.
  - Label: displayed only for nodes with three or more biclusters or an eigenvector centrality value equal to or greater than 0.3.
- **Edges.**
  - Color: color of one of the two linked nodes.
  - Size: arbitrary and uniform.
- **Layout: ForceAtlas** (home-brew layout of Gephi).

Appendix B.8 provides some insights about the significance of detecting communities and the Louvain algorithm.

2.10 Representation of SNPs in Gene Maps.

mapsnp v2 [60] is an R package that accesses the **UCSC database** (`https://genome.ucsc.edu/`) and allows for the creation of gene maps, representing the desired SNPs within the gene. We used this package to create a **gene map for each affected gene within each bicluster**, displaying only the **SNPs** belonging to that bicluster. To achieve this, the first step was to create a table of SNPs grouped by gene for each bicluster. Each of these tables was iterated through by a *for* loop, which gathered the **data required by the mapsnp function *msb*** to create each gene map. This required data includes:

- Initial and final **genomic positions of the gene** in the reference genome GrCh37 for Homo sapiens, obtained using the R BiomaRt package.
- Data obtained from Ensembl:
  - **Identifiers** for SNPs within the gene and the bicluster, which were also merged with SNP **"Impact"** values to provide additional insights.
  - **Genomic coordinates** for each SNP.

  - **Chromosome** name where the SNPs (and the gene) are located.
- **Gene name**: we translated the Ensembl identifier to the external name (using the KnowSeq package) and included it in the *msb* query.

2.11 Bicluster-Specific Analyses.

As previously explained, most of the data analysis tools mentioned earlier can be applied to the entire set of patients or SNPs, or exclusively to those specific to a bicluster.

For the bicluster-specific analysis, first we created a **comprehensive bicluster summary table**. In this table, each row corresponds to a bicluster, except for the first row, which references the total data (unique patients/SNPs/genes) studied in this project. The columns represent numeric attributes related to patient clinical information, SNP types, and types of genes affected by these SNPs. To enhance the visualization of this table, we utilized the **gt package** in R.

Then, we constructed a function that takes only the bicluster name as input and provides the following results:

- The information related to the bicluster appearing in the bicluster summary table is printed.
- If present, bicluster's **intergenic SNPs** with a probability of 0.25 or higher of having regulatory function in the organs of interest (according to RegulomeDB) are displayed.
- For each of these intergenic SNPs, the **genotypic percentages and the minor allele frequency** [20] in the bicluster are printed.
- An **ORA** applied to all affected genes in the bicluster is executed (with the same characteristics as those described in subsection 2.7).
- **For each affected gene in the bicluster:**
  - A **gene graph** (created with mapsnp) is displayed.
  - It is specified whether a **relationship** (direct or indirect) between the gene and **CRC or cancer in general** has been found. Information about these relationships (if any) and the general function of the gene is provided.
  - When the gene belongs to one of those identified as **DEGs** between cancerous colon and healthy colon, this information is reported, including the DEG statistics.

[20] The minor allele frequency was calculated as the sum of genotype 3 frequency and half of the frequency of genotype 2, representing the frequency by chromosome. The concept of hemizygosity was taken into account to adjust the calculation of the minor allele frequency.

  - For each of the **SNPs** appearing in the gene, the percentages of **genotypes** in the corresponding bicluster are shown, as well as the **minor allele frequency**.
  - A table displays some **Ensemble information** about each of these SNPs (consequence type, impact and sift prediction [21]).
- Finally, as a **summary of the bicluster SNPs**, **two tables** are generated with the gt package. One for intergenic SNPs (only those filtered based on the probability of being regulatory in the organs of interest), and another for all genic SNPs. Both tables include genotypic information related to the bicluster. The intergenic SNPs table also includes RegulomeDB attributes while the genic SNPs table includes Ensembl attributes. Additionally, for genic SNPs, information about the gene (name, whether it has a direct/indirect relationship with CRC/cancer in general, and whether it has been detected as a DEG) is also included.

For each bicluster, in addition to executing this function, we also created **bar charts of the bicluster-specific clinical table.**

2.12 Identification of Interesting Biclusters and Genes.

The set of operations described in the previous subsection produced a significant amount of information that allowed to identify biclusters that appeared more interesting than others. Specifically, **two particularly interesting biclusters were selected** for which **additional analytical tests** were conducted, such as expanding the annotation of their affected genes or constructing protein-protein interaction graphs with them.

Similar to the identification of interesting biclusters, the analytical operations explained in Chapter 2 allowed for the identification of some **genes that appear particularly intriguing**. Their interest arises from the fact that one or several of the **extracted results suggest that their SNPs were especially relevant in the identification of PGMRA CRC-biclusters**. We paid special attention to genes appearing in a higher number of biclusters or associated with a greater number of SNPs, also investigating whether external evidence (such as literature or results from our RNA-Seq data analysis) links them to CRC. Results for each gene, including gene maps or functional annotations, will only be presented for those identified as particularly interesting in this regard.

[21] Prediction of variation effect on protein function.

2.13 Genotypic Cluster-Heatmap.

We identified **two biclusters that collectively represent a substantial portion of the 73 CRC patients**, prompting us to find it interesting to analyze and compare their **genotypic data** using cluster-heatmaps. To achieve this, we first generated a dataframe containing genotypic data for SNPs and patients in at least one of those biclusters. **Only the genic SNPs were selected**, and their names were transformed into a longer version that includes a series of attributes of interest provided by the genic SNP Ensembl dataset.

For each SNP/patient in these biclusters, information was gathered about their **dual or singular bicluster membership** (and if singular, which one). This information was added as annotations on the cluster-heatmaps. **Patient sex** details were also included to distinguish patterns due to sex rather than bicluster affiliation. To create heatmaps loaded with this amount of information, we decided to use the **ComplexHeatmap** package from R, Bioconductor. This toolkit enabled us to experiment with various methods for performing hierarchical clustering on rows and columns of the cluster-heatmap. It also allowed us to create heatmaps without applying clustering and explore different visualization options.

CHAPTER 3: RESULTS AND DISCUSSION

3.1 Missing Values.

In this section, the results of the missing values of genotypic and clinical data are presented. These analyses were conducted **after removing the four erroneous samples** and before any further preprocessing steps.

Genotypic Missing Values.

The genotypic table consisted of 73 rows (patients) and 2496 columns (SNPs). The **overall percentage of missing values** in this dataset was **0.046%**. The highest percentage of missing values for a patient was 1.20%, and for an SNP, it was 1.37%. The number of patients with at least one missing value was 30, and the number of SNPs with at least one missing value was 84. This implied that a significant proportion of patients had at least one missing value, but these missing values were concentrated in a small proportion of SNPs.

Clinical Missing Values.

The clinical dataset consisted of 73 patients (rows) and 87 columns that corresponded to clinical attributes and bicluster membership variables. This dataset exhibited a **much higher overall percentage of missing values** compared to the previous one, specifically 21.17%. Among the clinical attributes, 43 had at least one missing value, and 14 were completely empty. All columns with more than 90% of missing values were removed. Regarding patients, one instance exhibited 49.42% of missing values.

3.2 Hemizygosity Encoding.

This section displays the results of analyzing genotypic subdatasets to explore hemizygosity (presence of a single allele in an individual) encoding.

- Genotypic subdataset of **male patients and X chromosome SNPs** (56 patients and 574 SNPs): in this dataset, the vast majority of genotypes are *1* or *3*, with only 0.84% being *2*. We verified that genotype 2 can appear in any of the patients but is accumulated in only 8 SNPs.
- Genotypic subdataset of **male patients and Y chromosome SNPs** (56 patients and 6

SNPs): most of the genotypes are *2*, with 6.54% being *3* and 0.29% being *1*.

- Genotypic dataset of **female patients and Y chromosome SNPs** (16 patients and 6 SNPs): most of the genotypes are *1*, except for 16.67% being *2* and 3.125% being *3*.

Hypothesis.

The fact that men mostly have genotypes *1* and *3* in X chromosome SNPs suggests that **hemizygosity in encoded in the same way as homozygosity**: if a male individual has the minor allele on their single X chromosome, this genotype is encoded as *1*, just as it is for women who have the minor allele on both of their X chromosomes.

However, there are **8 X chromosome SNPs that do not follow this rule**, as men can be heterozygous (genotype *2*) for them. Similarly, for the **6 Y chromosome SNPs**, both men and women can be heterozygous (although men have only one Y chromosome and women have none). Therefore, it seems that both the X chromosome SNPs for which men can be heterozygous and the 6 Y chromosome SNPs are actually located on both sex chromosomes, i.e., in **pseudoautosomal regions**, causing each individual to have **double dose** of these SNPs regardless of their sex.

Confirmation of the hypothesis.

To investigate this, we checked the chromosome on which these 14 SNPs were located according to the **bim file**, which is a file produced by the Plink software, following Illumina sequencing. We found that all of them had **"X/Y" as their chromosomal category,** which was a discrepancy with Ensembl. In fact, aside from these SNPs, there were no additional SNPs associated with "X/Y" (except for some SNPs with identifiers not recognized by Ensembl, with which we did not work). This finding confirms that hemizygosity is encoded by genotypes *1* and *3*, and sex chromosome SNPs that may appear with genotype *2* are found in pseudoautosomal regions.

Consequences of this finding in the project.

We did not change the chromosome names of pseudoautosomal SNPs to "X/Y" in the Ensembl annotation tables because some of these SNPs are associated with genes linked to only one of the sex chromosomes (even if there is a homolog). [22] However, we took into account that these

[22] For example, rs2524578, one of the SNPs in pseudoautosomal regions, according to Ensembl, is located on chromosome X, associated with the gene PCDH11X (protocadherin 11 X-linked), which has its homolog on the Y chromosome, PCDH11Y [61].

SNPs are present in double dose, to treat them the same as SNPs on autosomes when calculating the **frequency of the minor allele**.

3.3 Analysis of Similarity Among PGMRA Biclusters.

3.3.1 Patient Sharing Among PGMRA Biclusters.

Figure 8 corresponds to the **percentage matching matrix** of patients between biclusters, with the number of patients in each bicluster along the main diagonal. The biclusters with the **highest number of patients** are 8.4 (54 patients), 2.2 (35), 11.5 (35), and 10.1 (34). Among the latter three, the bicluster with which 8.4 shares the lowest percentage of patients is 2.2 (50% of 8.4 patients are included in 2.2). Thus, the pair of biclusters that includes the highest proportion of patients is formed by **8.4 and 2.2**. Specifically, both biclusters account for **58 out of 73 patients** (almost 80%).

Considering that bicluster 8.4 has a significantly higher number of patients than the rest, it was expected that a large proportion of patients from most groups would also be included in this bicluster. Therefore, the column for 8.4 has the highest values in the matrix. Specifically, 10 out of the remaining 15 biclusters have more than 85% of their patients shared with bicluster 8.4. Thus, **bicluster 8.4. seems to capture the general SNP commonalities shared by most CRC patients**.

On the other hand, we observe **particularly small groups**, such as 3.2 and 10.4, each with only 3 patients. The latter is the only bicluster that does not share any patient with 8.4, the "general" bicluster.

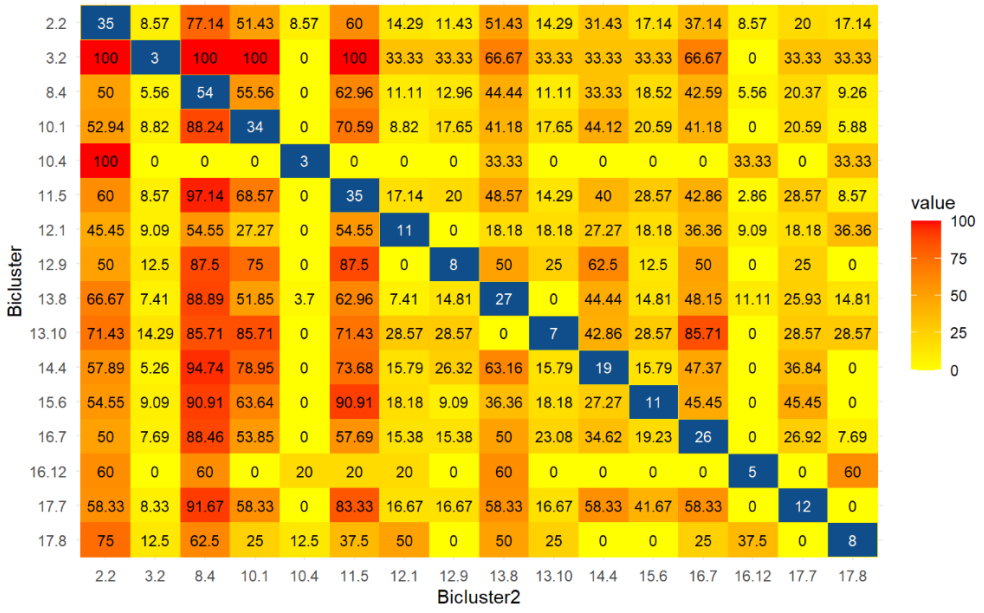


Figure 8. Percentage matching matrix of patients between biclusters. Each value indicates the percentage of patients from the row's bicluster that are also included in the column's bicluster. Diagonal values represent the number of patients from the bicluster that appear both in the corresponding row and column.

Figure 9 illustrates the **weighted directed graph** derived from the earlier matrix. The nodes with the **highest eigenvector centrality** are 2.2, 8.4, and 11.5 (distinguished by their darker color). However, the ranking differs for **betweenness**, where, as indicated by the order of appearance in the clockwise direction, the sequence is 2.2, 13.8, 17.8, followed by 8.4 and 11.5. These outcomes suggest that, despite not having the largest number of patients (35), **2.2 is the most crucial node** in the graph, based on its node centrality concerning linked nodes and the number of minimum paths passing through it. This result was expected, as bicluster 2.2 corresponds to the execution of the fuzzy NMF algorithm with the number of biclusters set to 2.

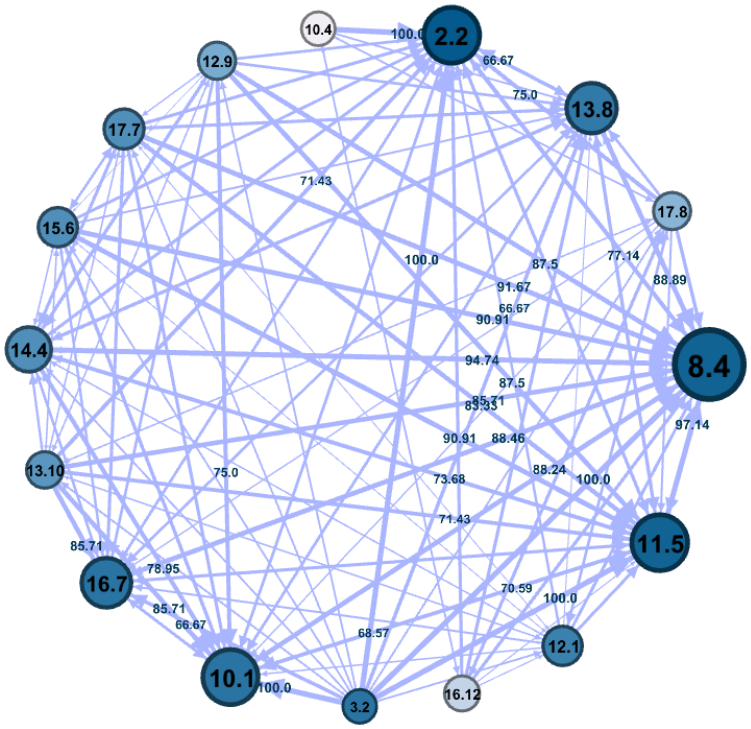


Figure 9. Bicluster patients directed weighted graph. Node color: based on eigenvector centrality (darker indicates higher centrality). Node size: based on the number of patients (larger represents more patients). Edge color: arbitrary. Edge size (including edge label size): based on weight (percentage of patients from the source node also present in the target node). Edge label: only displayed for weights (edge labels) greater than 65. Layout: circular, with nodes arranged in a clockwise direction based on decreasing betweenness centrality values.

3.3.2 SNP Sharing Among PGMRA Biclusters.

In a similar fashion to Figure 9 with patients, Figure 11 illustrates the **SNP matching matrix** between biclusters.

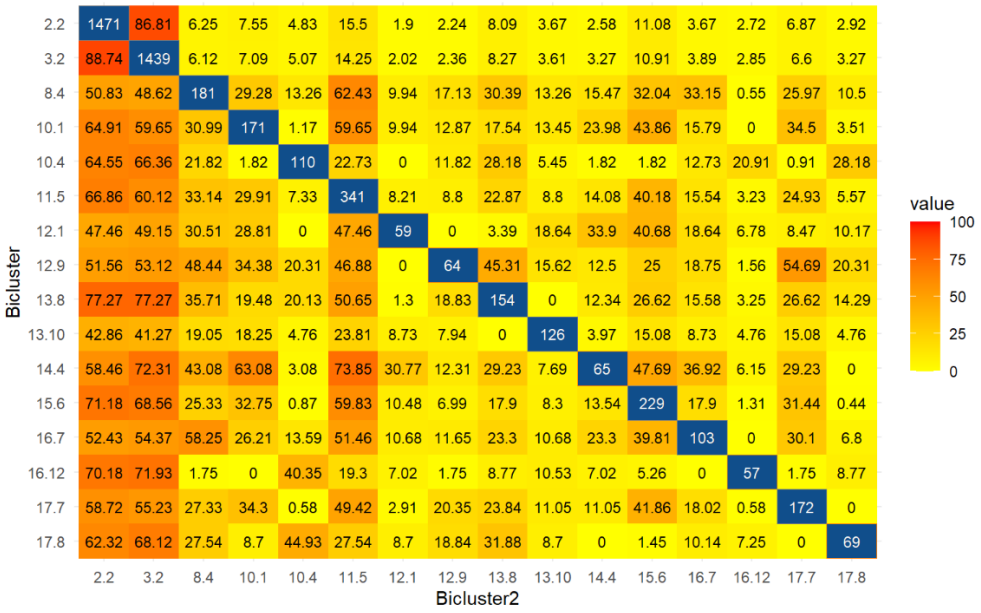


Figure 10. Percentage matching matrix of SNPs between biclusters. Each value indicates the percentage of SNPs from the row's bicluster that are also included in the column's bicluster. Diagonal values represent the number of SNPs from the bicluster that appear both in the corresponding row and column.

Biclusters 2.2 and 3.2 have significantly **more SNPs** (1471 and 1439, respectively) than the rest. Remarkably, they share over 85% of their SNPs with each other. Bicluster 13.10 stands out with the lowest percentage of SNPs included in 2.2 (42.86%) and 3.2 (41.27%). **Bicluster 8.4**, despite having the most patients, has **relatively few SNPs** (181). A larger percentage of these SNPs is shared with bicluster 11.5 (62.43%) compared to 2.2 (50.83%) or 3.2 (48.62%). Finally, the biclusters with the **smallest number of SNPs** are **16.12 and 12.1**.

Figure 11 corresponds to the **weighted directed graph derived from the SNP matching matrix**. The configuration of the topology and visualization of this graph is the same as explained for the patient matching graph.

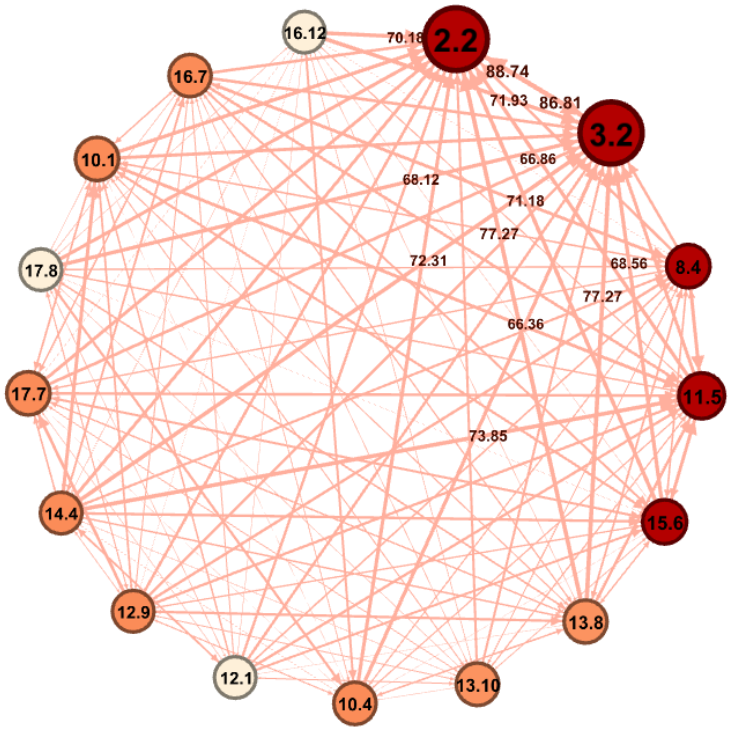


Figure 11. Bicluster SNPs directed weighted graph. Node color: based on eigenvector centrality (darker indicates higher centrality). Node size: based on the number of SNPs (larger represents more SNPs). Edge color: arbitrary. Edge size (including edge label size): based on weight (percentage of SNPs from the source node also present in the target node). Edge label: only displayed for weights (edge labels) greater than 65. Layout: circular, with nodes arranged in a clockwise direction based on decreasing betweenness values.

In this graph, the **top 5 ranking** of nodes with the highest **eigenvector and betweenness centralities coincides**. These nodes are 2.2, 3.2, 8.4, 11.5 and 15.6 and **they are tied** in both metrics. Therefore, similar to the patient matching graph, biclusters 2.2, 8.4 and 11.5 maintain high centrality. Biclusters 3.2 and 15.6 achieve greater centrality in the SNP graph than in the patient graph, likely due to the fact that, while they have a very low number of patients, they are the second and fourth biclusters, respectively, with the highest number of SNPs. 10.4 is another bicluster that stands out for having much higher centrality in the SNP graph than in the patient graph (where it was one of the least central biclusters).

3.4 Analysis of the Entire Sets of Patients and SNPs.

This section presents the results of the analyses of the entire sets of patients and SNPs found in any of the 16 CRC-biclusters, regardless of their specific bicluster memberships.

3.4.1 Patient Clinical Data Analysis: Bar Charts and Association Rules.

As we saw in Subsection 3.1, the clinical table has many columns with a high number of missing values, and others were considered redundant. Therefore, the preprocessing steps excluded several columns, leaving a total of 29 to work with, which are: "Age_Range", "Sex", "Metastasic", "Surgery", "Histolog", "Loc_T" (tumor locality), "Loc_T2" (a variable that groups the values of "Loc_T" into new categories), "Stage", "Degree", "Basal_CEA" (basal Carcino Embryonic Antigen, a tumoral marker), "Met_liver", "Met_lung", "Met_lymph_nodes" and the 16 columns corresponding to the biclusters.

Bar Charts of Clinical Attributes.

The 16 bicluster-related variables have binary values indicating membership. However, instead of creating a bar chart for each bicluster, figure 12 displays a single bar chart indicating the number of instances with a value of "1" in each of these columns, thus representing the **number of patients in each bicluster** (information already known from the patient matching matrix).

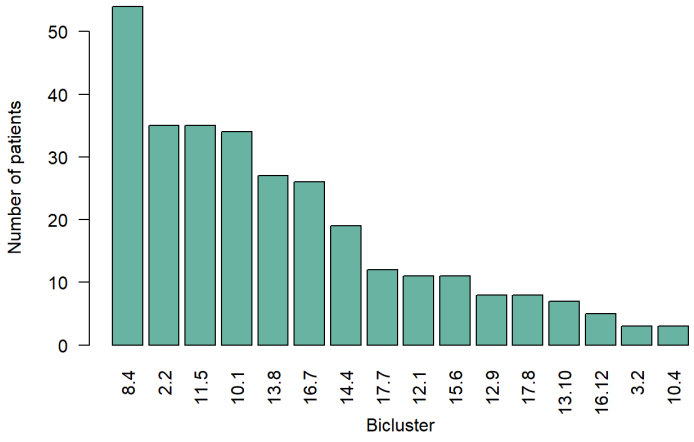


Figure 12. Bar chart representing the number of patients in each bicluster.

Figures 13 and 14 display a bar chart for each of the selected clinical attributes (charts are relative to the preprocessed dataset).

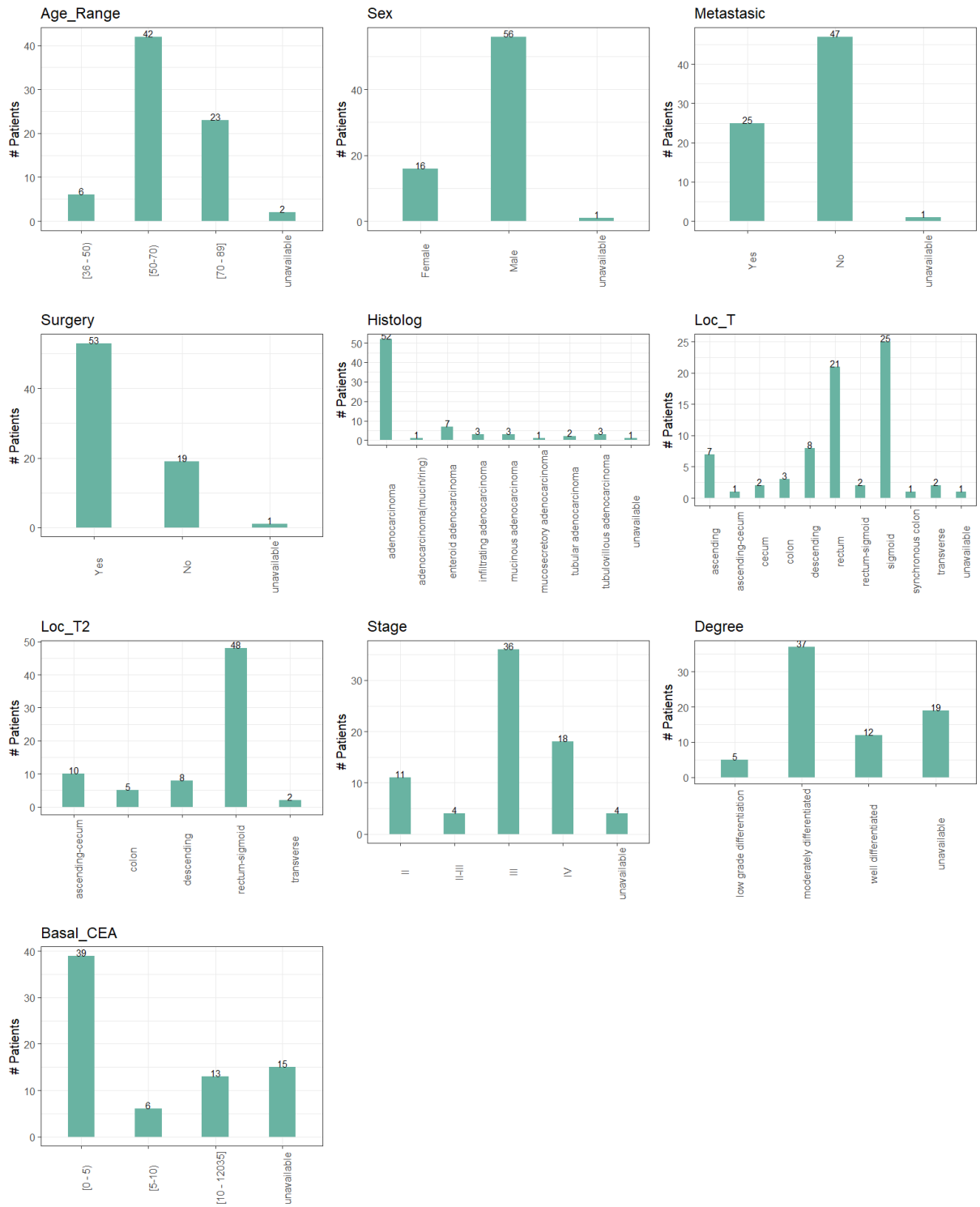


Figure 13. Bar charts of several clinical attributes: "Age_Range", "Sex", "Metastasis", "Surgery", "Histog", "Loc_T", Loc_T2", "Stage", "Degree" and "Basal_CEA".

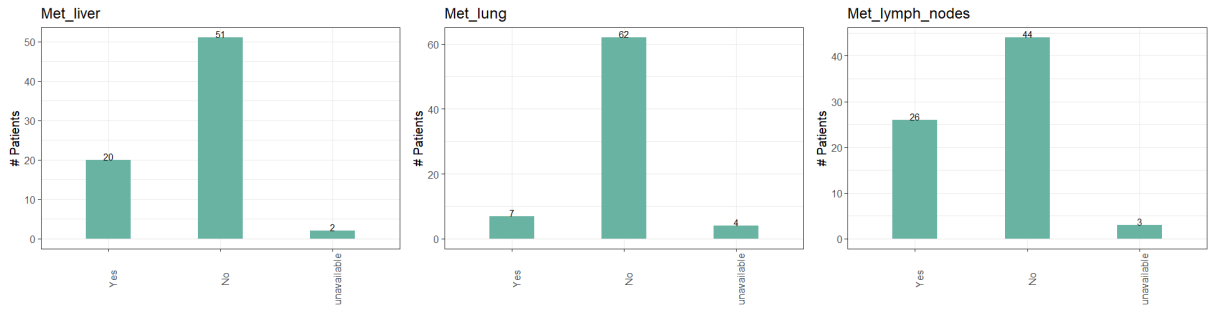


Figure 14. Bar charts of several clinical attributes: "Met_liver","Met_lung" and "Met_lymph_nodes".

From these figures, we can highlight the following points:

- The **age range** with the highest number of patients is [50-70), with only 6 patients below the age of 50.
- The number of **male patients** is significantly higher than that of females.
- There are considerably more **metastatic patients** than non-metastatic ones and more patients who underwent **surgery** than those without surgery.
- The most frequent **tumor locations** are "sigmoid" and "rectum", which are grouped into the same locality with the "Loc_T2" reclassification of "Loc_T".
- The most common **tumor stage** is, by far, Stage III, and the most frequent **degree of differentiation** is "moderately differentiated". Only 5 patients have "low-grade differentiation".
- More than half of the patients have a **"basal_CEA"** value between 0 and 4.
- The number of patients with **liver, lung, and lymph node metastasis** is 20, 7, and 26, respectively.

Extraction of Association Rules from the Clinical Table.

From the clinical table, **51,795 association rules** were extracted with a length between 2 and 4 items, a minimum support of 0.05, a minimum confidence of 0.7, and a minimum lift of 1.4. After **removing redundant rules** [23], the number of rules was reduced to **14686**. Here is a concise selection of those deemed to be more informative:

[23] A rule is considered redundant if it is equally or less predictive than a more general rule, which has the same items on the consequent, but one or more items less on the antecedent.

Table 2. Subset of the association rules that were generated with a length between 2 and 4 items, a minimum support threshold of 0.05, a minimum confidence threshold of 0.7, and a minimum lift threshold of 1.4.

rules	supp.	conf.	cover.	lift	count
{Degree=low grade differentiation} ⇒ {2.2=1}	0.055	0.8	0.068	1.669	4
{Loc_T=ascending} ⇒ {16.7=1}	0.068	0.714	0.096	2.005	5
{Loc_T=descending} ⇒ {2.2=1}	0.082	0.75	0.11	1.564	6
{16.12=0,17.8=1} ⇒ {Metastasic=Yes}	0.068	1	0.068	2.92	5
{Basal_CEA=(10 - 12035)} ⇒ {Metastasic=Yes}	0.151	0.846	0.178	2.471	11

Simple association rules explanation.

- Bicluster 2.2 identifies patients with low grade of differentiation.
- Bicluster 16.7 identifies patients with ascending tumor localization.
- Bicluster 2.2. identifies patients with descending tumor localization.
- The 5 patients who are in 17.8 without being in 16.12 have metastasis.
- 11 out of 13 patients with basal CEA greater than 10 have metastasis.

Global analysis of association rules.

Regarding **rules concerning items related to biclusters**, we observe that, in several cases, patients sharing the same value for a clinical attribute tend to belong to the same bicluster. For example, 5 out of 7 patients with "Loc_T=ascending" are in bicluster 16.7. However, the bicluster is not exclusive or does not have such a strong over-representation of that item (bicluster 16.7, apart from these 5 patients, has 21 other patients with different "Loc_T" values). The first three rules reflect similar cases.

In Table 2, the only rule indicating over-representation of an attribute in a subgroup of patients is the **fourth one**. This subgroup comprises patients from bicluster 17.8 who are not included in 16.12. There are 5 patients in this subgroup, all sharing the presence of **metastasis**—an item that the majority of patients in the total dataset lack, resulting in a high lift for this rule.

Regarding the **last rule**, which is not related to biclusters, it was expected that, given **CEA's**

role as a tumor marker, its value would be higher in metastatic cases, as these indicate advanced tumor states. However, the rule informs us that even a baseline CEA level above 10 already correlates with the presence of metastasis in this sample of 73 CRC patients. We found in the literature that over-expression of CEA is closely associated with metastasis in the liver, which is the main cause of death from CRC [62].

3.4.2 Genic SNPs Analysis.

The total number of **SNPs** found in any of the 16 CRC-biclusters is 2,496, of which 972 are **genic**, 1,025 are **intergenic**, and 499 were **not recognized by Ensembl**, so they were excluded from the analysis. This subsection presents the analysis of the genic SNPs, which affect a total of 695 different **genes**.

Genic SNP Ensembl Table.

The most relevant attributes in the genic SNP Ensembl table were selected to construct **bar charts**, and all of them are presented in Figure 15.

Although this table includes Ensembl attributes related to the minor allele (possible alleles, minor allele, minor allele frequency), these attributes were discarded because the minor allele does not correspond, in most cases, to what has been considered the minor allele for encoding genotypes (*1*, *2*, and *3*) in our data (the causes of which are unknown). Consequently, we do not know if the **consequence type of the transcript variant**, the **impact** and the **SIFT prediction** of each SNP (variables with which we do work) are associated with the minor or major allele in our data. We will simply interpret these attributes as **the most severe consequence** that can result from changing one nucleotide to another at the **SNP genomic coordinate**.

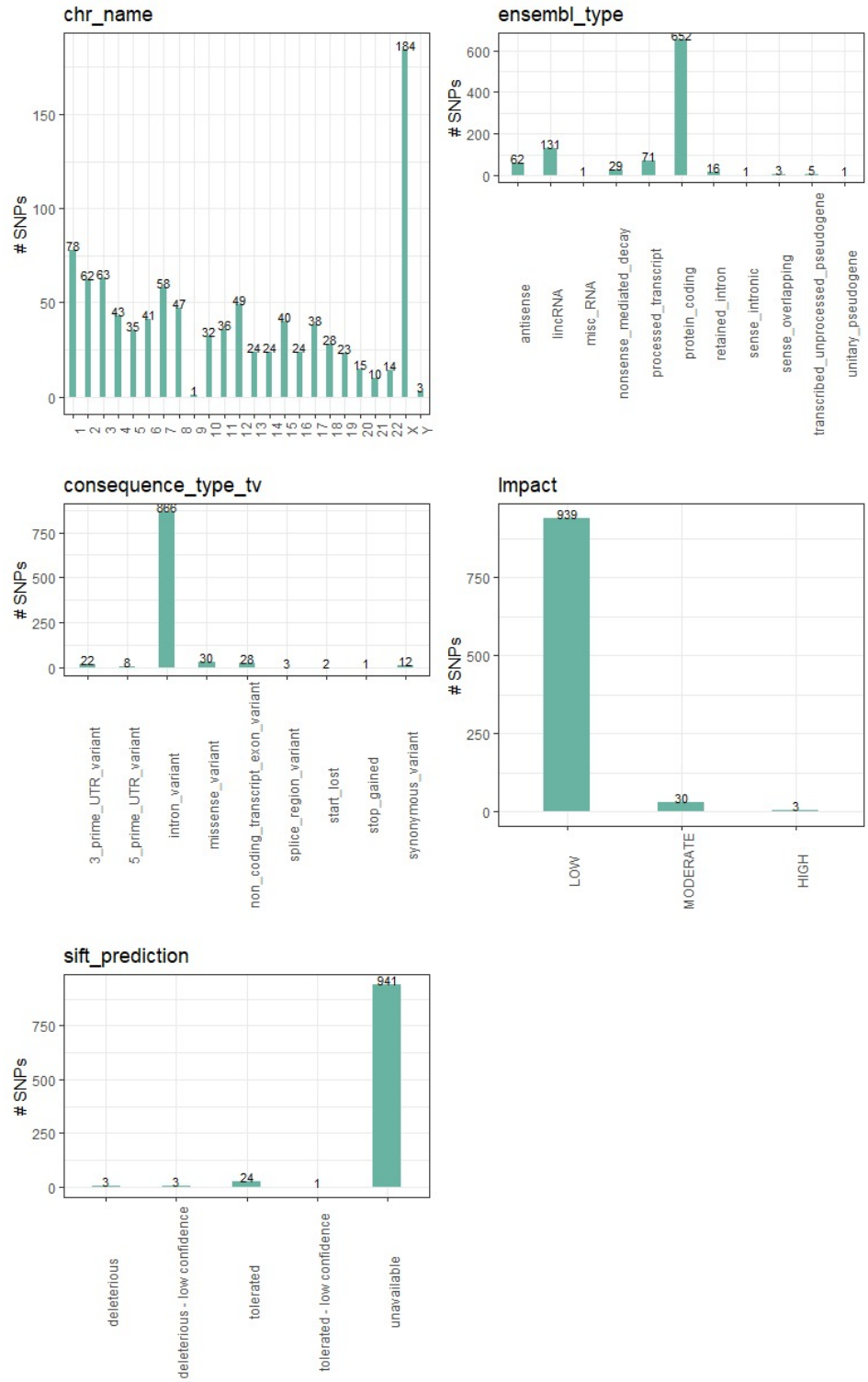


Figure 15. Bar charts of key attributes in the genic SNP Ensembl table: "chr_name", "ensembl_type", "consequence_type_tv", "Impact" and "sift_prediction".

From this figure, we can highlight the following points:

- The **X chromosome** significantly contains the highest number of genic SNPs (184). The next most frequent chromosome is **chromosome 1**, with 78 SNPs. On the other hand, **chromosomes 9 and Y** stand out for their low number of genic SNPs, with only 1 and 3, respectively.
- The majority of genic SNPs, specifically 652, are located in **protein-coding genes**. The next most SNP-accumulated gene type is the one encoding long non-coding RNAs (**lncRNAs**), with 131 SNPs.
- Out of the 972 genic SNPs, 868 produce **intronic variants** at the RNA transcript level. Since this variant is associated with a low impact level, it makes sense that the most frequent impact value is overwhelmingly **"LOW"**.

The low presence of SNPs causing a severe change is not surprising. As mentioned in the introduction, current research indicates that the genetic variation of numerous traits is predominantly shaped by multiple regions in the genome, each exerting a small influence.

Additionally, we wanted to **analyze if SNP severity had a differential frequency based on the chromosome**. Figures 43 and 44 illustrate the distribution of the SNP consequence type and impact as a function of the chromosome name. Both graphs and their overview can be found in Appendix C.1. However, from them, we can highlight that chromosome 1 accumulates more SNPs with **moderate impact** than chromosome X despite having many fewer genic SNPs.

Annotation of the Complete Set of Affected Genes.

The **annotation of the 695 genes** using VarElect, GeneCards, and the Bio package from Entrez (in Python), allowed us to identify direct relationships with CRC for 146 genes and indirect relationships with CRC for 416 genes. Moreover, 49 genes not related to CRC were related to cancer in general. For the remaining 86 genes, no connections were found with either colon cancer or cancer in general. Table 3 summarizes this information.

Table 3. Number of genes for which a relationship (direct or indirect) with CRC or cancer in general has been found.

Affected Genes	Number of genes	Percentage
Colon Cancer Directly Related	146	21.01
Colon Cancer Indirectly Related	414	59.57
Cancer Directly Related	25	3.6
Cancer Indirectly Related	24	3.45
No relationship found	86	12.37

Over-Representation Analysis of the Complete Set of Affected Genes.

First, the ORA was conducted, as explained in Subsection 2.7, considering the input of **695 affected genes**. However, the effective analysis could only be performed on genes with Entrez identifiers, which amounts to 581 out of the 695 genes.

The **15 most over-represented diseases** in this gene set are displayed in the bar graph in Figure 16. For each disease, we can observe the number of genes associated with it (x-axis value) and the approximate adjusted p-value (based on the bar color), indicating how significant the over-representation of that disease is in the query gene set. In this case, the size of the query gene set was 526, as the remaining genes were not recognized by the DisGeNET database. We can see that the majority of diseases are related to **mental disorders or neurological problems** and especially, there is a high presence of terms related to **drug addiction**.

This outcome is interesting and aligns with the **findings of other authors**. For instance, a study conducted by Shane Lloyd *et al.* [63] demonstrated an increased risk of various mental health disorders, such as depressive and cognitive disorders, as well as substance abuse, among CRC survivors compared to the general population cohort. Furthermore, it revealed that CRC patients who develop mental health disorders also experience decreased survival.

On one hand, the prevalence of these disorders in our gene set lends **external support to the reliability of PGMRA**. This is evident for two reasons: 1. PGMRA unbiasedly formed groups of patients exclusively diagnosed with colon cancer (as opposed to breast or skin cancer) and 2. Notably, this grouping is rooted in genetic variants (SNPs) significantly located in genes whose altered function links with conditions highly prevalent in CRC patients. On the other

hand, **our results could complement those of Shane Lloyd *et al.* study**, as these authors relate environmental factors (such as colostomy) to the development of mental illnesses in CRC patients, while we would be providing the **genetic aspect**—the apparent predisposition that certain SNPs confer to alter neurological function and contribute to the onset of colon cancer simultaneously. Of course, much more investigation is needed to confirm whether our SNPs are indeed causing these genes' function to be altered and eventually triggering these mental health disorders. Additionally, the role of genes related to mental illnesses in colon cancer remains unknown.

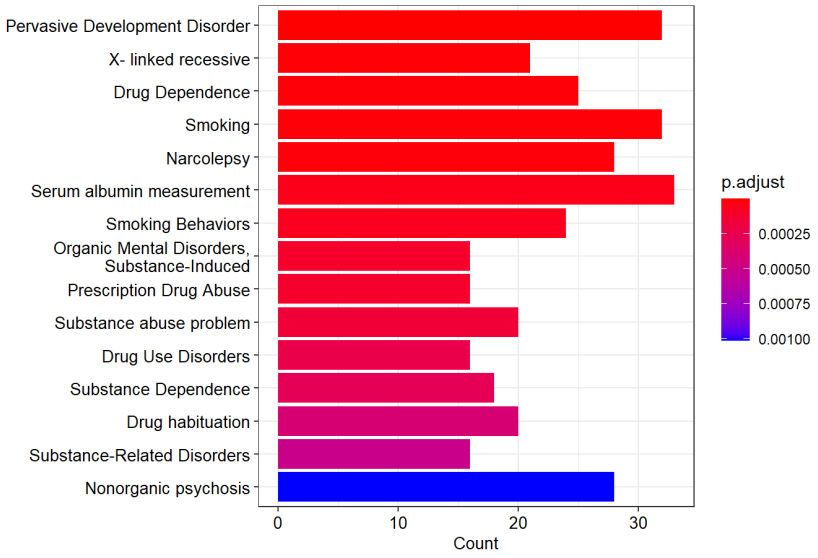


Figure 16. Bar chart of the 15 most over-represented diseases (with lowest adjusted p-value) in the total of affected genes. The size of the recognized query gene set in the DisGeNET database was 526.

It is important to note that while these are over-represented diseases in the total set of affected genes, it does not imply that there are no **many other relevant gene functions** related to colon cancer in this gene set. In fact, in this bar graph we can observe that there are not more than 35 genes associated with any of these diseases, when the total number of genes is 695. [24]

Table 4 presents the only **over-represented GO CC term** in the total set of genes (only 486 genes from the total set were identified by the ontology, as indicated by the "GeneRatio" field).

[24] What these over-representation results indicate is that the number of genes associated with any of these 15 diseases is significantly higher than what would be expected by chance in a gene set of size 526.

This term is **"neuron to neuron synapse"**, which aligns with the neurological issues observed in the diseases graph. In fact, over-representation of the neuron-to-neuron synapse function could be expected not only due to the prevalence of mental disorders in CRC patients but also because of the **prevalence of peripheral neuropathy** in this disease. Peripheral neuropathy is a detrimental and persistent complaint often attributed to chemotherapy. However, Boyette-Davis *et al.* study [64] showed that CRC patients exhibited subclinical deficits (for example in sensorimotor function or touch detection), before the exposure to this environmental factor.

Studies regarding the role of the nervous system in CRC remain limited, even though **the gastrointestinal tract is highly innervated**, both from outside the intestines (extrinsic innervation) and by a nervous system of their own; the enteric nervous system (intrinsic innervation) [65].

Table 4. Over-represented Cellular Component GO terms in the total set of affected genes.

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
GO:0098984	neuron to neuron synapse	28/486	7.44E-05	3.13E-02	2.60E-02

Over-Representation Analysis of the Set of Genes Directly Related to CRC.

The **ORA was conducted again**, using the same parameter configuration, for the **set of genes directly associated with colon cancer**, totaling 146 genes. Table 17 presents the 15 most over-represented **diseases** in this gene set. We can still observe some **mental disorders**, such as impulsive behavior, but there are also many **terms related to cancer** (childhood kidney Wilms tumor, gallbladder carcinoma, etc.), although not specifically to colon cancer. The appearance of diseases related to other cancers is not surprising, as many genes play a role in various types of cancer or neoplasms (abnormal cell growth) in general [66].

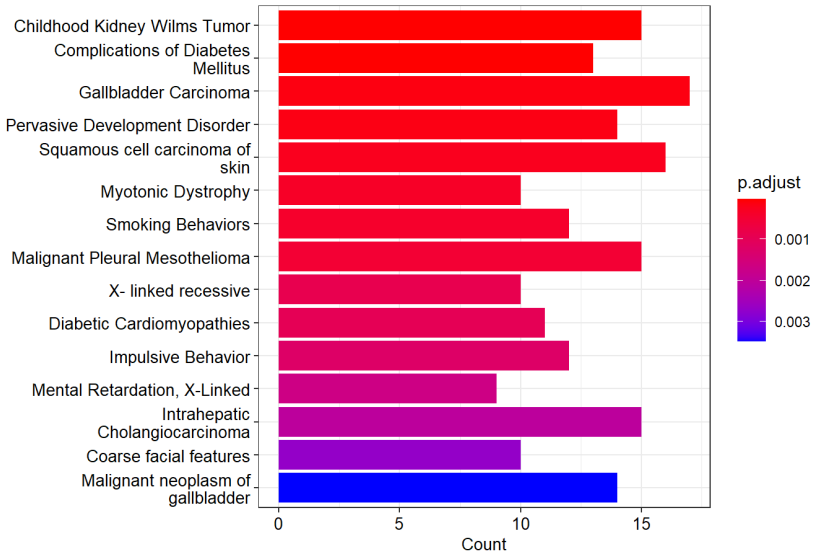


Figure 17. Bar chart of the 15 most over-represented diseases (with lowest adjusted p-value) in the genes for which a direct relationship with CRC has been found. The size of the recognized query gene set in the DisGeNET database was 141.

Table 5 displays the **over-represented CC GO terms** in the genes directly associated with CRC. In this case, the term "neuron-to-neuron synapse" that appeared for the 695 genes is not present, but most of the terms that appear here are similarly **related to synapses**. This outcome suggests that indeed a significant proportion of the genes we are working with are simultaneously associated in the literature with CRC and the synaptic function.

Table 5. Over-represented Cellular Component GO terms in the set of genes directly related to CRC.

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
GO:0098793	presynapse	19/133	1.21E-06	3.53E-04	3.04E-04
GO:0044306	neuron projection terminus	9/133	3.73E-06	1.09E-03	4.68E-04
GO:0099522	cytosolic region	4/133	1.92E-05	5.60E-03	1.60E-03
GO:0043679	axon terminus	7/133	9.66E-05	2.81E-02	6.05E-03
GO:0099524	postsynaptic cytosol	3/133	1.59E-04	4.64E-02	7.98E-03

We also found over-represented **pathways** in the set of genes directly related to CRC, as shown in Table 6. It is interesting the presence of the **Calcium Reabsorption pathway**, given that

the intestines can absorb considerable amounts of calcium via Ca2+ permeable ion channels and **hypercalcemia is common in patients with CRC**, who often present abnormal calcium channel protein expression, as stated by Wang *et al.* [67]. The study conducted by these authors revealed that even a slight upregulation of intracellular Ca2+ signaling can facilitate the onset and progression of CRC. Furthermore, the **remaining pathways** shown in Table 6 are also related to **Ca2+**, especially the second and third ones:

- **The Phosphatidylinositol Signaling pathway** leads to the release of Ca2+ from the endoplasmic reticulum [25], stimulating the occurrence of the Ca2+ signaling pathway that mediates many cellular life activities such as inflammation, metabolism, apoptosis [26], etc. [70]
- The third pathway refers to the synthesis and function of the **Parathyroid Hormone (PTH)**, which precisely controls calcium levels in the blood.
- Finally, **focal adhesions** are sub-cellular structures that mediate the regulatory effects, such as signaling events, of a cell in response to extracellular matrix adhesion. This pathway may be related to the previous ones, as the Phosphatidylinositol Signaling pathway is activated by the binding of an extracellular signal molecule to a receptor on the cell surface.

These results, once again, seem to **externally validate the PGMRA algorithm** while pointing towards the **extraction of novel knowledge**: it appears that the hypercalcemia observed in CRC patients could have its roots in some of the genic SNPs on which PGMRA has relied to group these patients.

[25] The endoplasmic reticulum is the largest organelle in the cell and is a major site of calcium storage, among other functions [68].

[26] Apoptosis is a tightly regulated cell suicide program that plays an essential role in the maintenance of tissue homeostasis by eliminating unnecessary or harmful cells. The evasion of apoptosis has been recognized as a hallmark of cancer [69].

Table 6. Over-represented pathways in the set of genes directly related to CRC.

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
hsa04961	Endocrine and other factor-regulated calcium reabsorption	7/82	5.88E-07	1.29E-04	9.84E-05
hsa04070	Phosphatidylinositol signaling system	7/82	3.50E-05	7.71E-03	2.93E-03
hsa04928	Parathyroid hormone synthesis, secretion and action	7/82	6.20E-05	1.37E-02	3.46E-03
hsa04510	Focal adhesion	9/82	1.23E-04	2.71E-02	5.16E-03

Gene Expression Analysis Using External RNA-Seq Data.

The analysis of gene expression in healthy and cancerous colorectal tissue using real RNA-Seq data from TCGA identified **3204 DEGs** in the two types of samples, with a minimum logFC of 1.5 and a maximum p-value of 0.001. Out of the 695 genes investigated in this master thesis, **72** belonged to this set of DEGs. Specifically, 43 were downregulated, and 29 were upregulated.

Figure 18 displays boxplots of expression in each sample type for a subset of 50 out of the 72 DEGs, chosen for better visualization. For some genes, the difference in mean expression between both samples is more pronounced than for others. Additionally, some genes exhibit expression ranges that overlap in both sample types when considering the outliers.

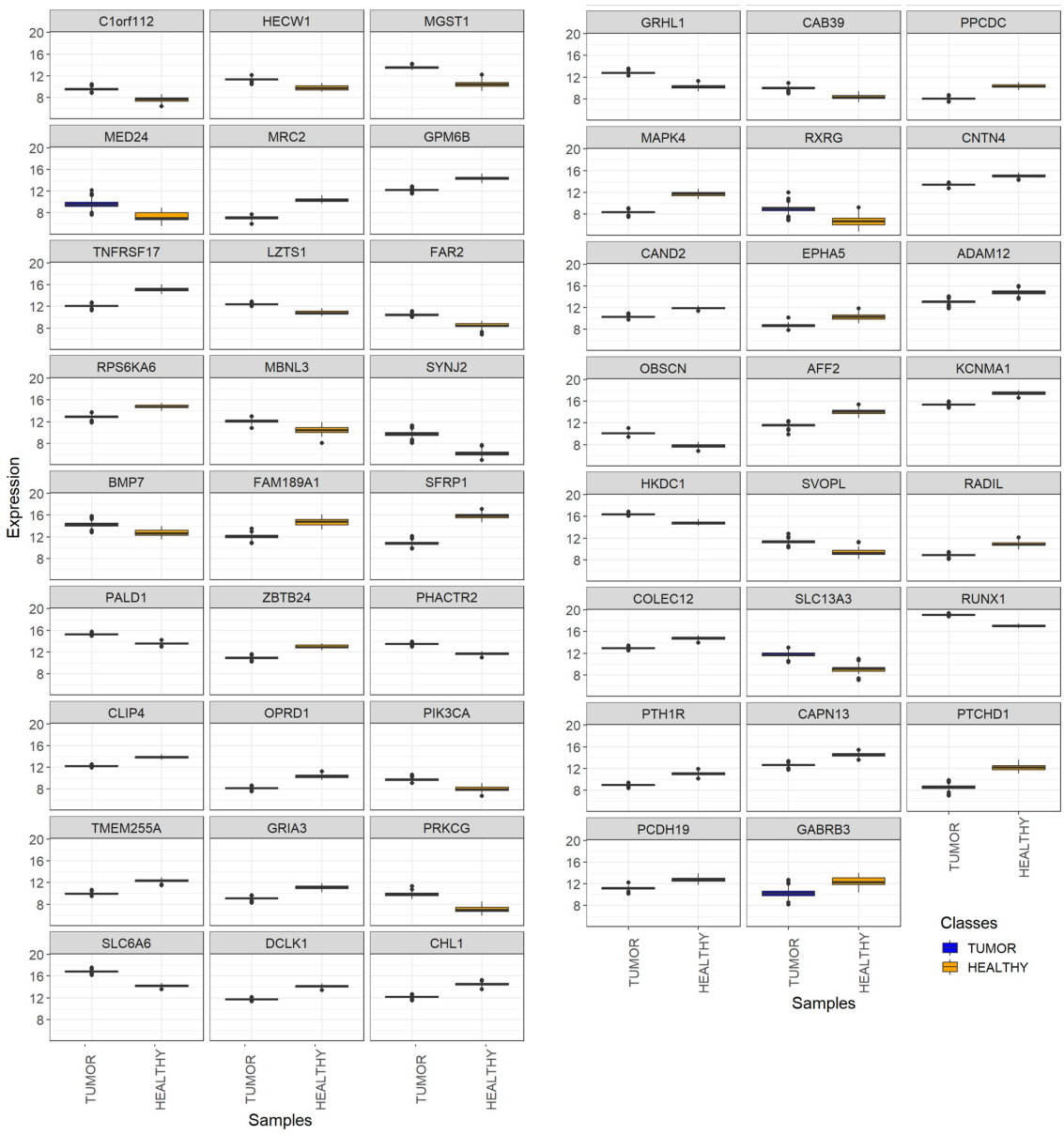


Figure 18. 50 out of the 72 genes identified as DEGs (from the total of affected genes) in colorectal cancer vs. healthy colon/rectum: boxplots of their expression in the two sample types, after preprocessing of TCGA data.

Figure 19 displays a heatmap of the expression in each sample for the 72 DEGs. For all of them, we can observe the difference in colors in both sample types, indicating differential expression.

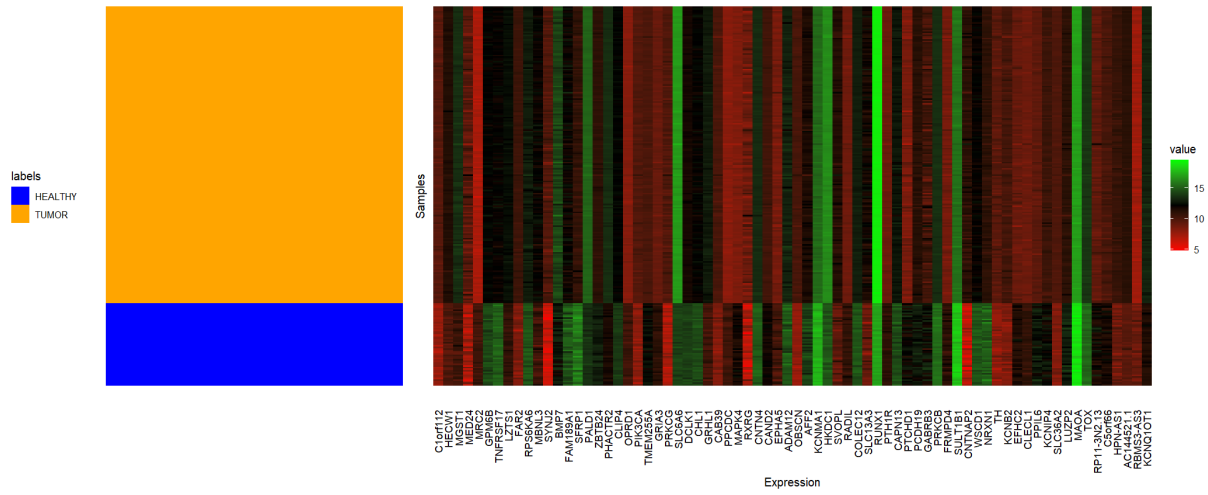


Figure 19. 72 genes (from the total of affected genes) identified as DEGs in colorectal cancer vs. healthy colon/rectum: heatmap of their expression in the two sample types, after preprocessing of TCGA data.

We cannot assert that the SNPs located in these 72 genes are altering their expression, which would classify them as eQTLs (Expression Quantitative Trait Loci, genomic regions associated with variation in gene expression levels [71, 72]). Nevertheless, the altered expression of these genes in CRC suggests that they play a significant role in the onset or progression of the disease. This role could be influenced by the presence of SNPs in any manner.

For this set of 72 DEGs, Table 7 displays the percentage of genes for which a relationship with CRC or cancer in general has been found. Additionally, the percentages relative to the total set of 695 genes are included for ease of comparison. The table allows for the identification of an **increase in the percentage of genes related to CRC (directly or indirectly) in the DEGs** group (98.61%) compared to that observed in the total set of genes (80.58%). These associations of the identified DEGs with CRC in the literature reinforces the validity of the RNA-Seq analysis carried out.

Table 7. Percentage of genes for which a relationship (direct or indirect) with CRC or cancer in general has been found: a comparison in the total set of affected genes and those identified as DEGs. ("D.CRC" ≡ Directly related to CRC, 'IND.CRC" ≡ Indirectly related to CRC, "D.C" ≡ Directly related to cancer, "IND.C" ≡ Indirectly related to cancer, "NO REL" ≡ No relationship found with CRC nor cancer in general).

-	%D.CRC	%IND.CRC	%D.C	%IND.C	%NO.REL
All Genes	21.01	59.57	3.6	3.45	12.37
DEGs	31.94	66.67	1.38	0	0

Protein-Protein Interaction Network Derived from the Total Set of Affected Genes.

Figure 20 displays the **protein-protein interaction minimum network** created using NetworkAnalyst (processed with Cytoscape and Gephi), taking the **695 genes as input**. Among these, **398** are integrated into the network as **seed nodes**, highlighted in a reddish tone. This reddish tone is darker for nodes that appear in a **higher number of PGMRA biclusters**. The remaining nodes, colored in blue and totaling 452, connect seed nodes.

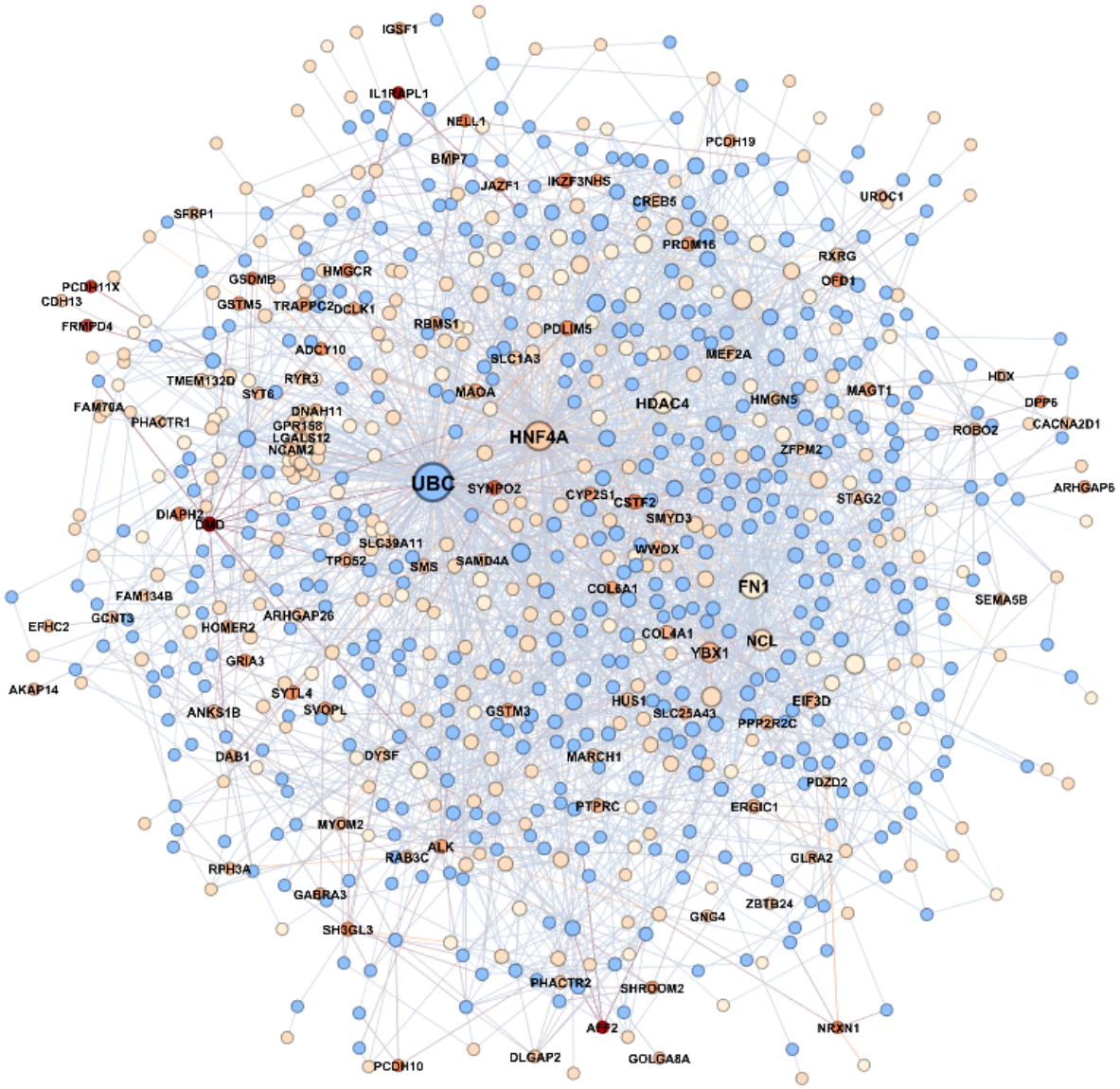


Figure 20. Minimum protein-protein interaction network derived from the total set of affected genes. Interactions were obtained using NetworkAnalyst, and visualization was done with Gephi. The blue nodes represent non-seed nodes (not part of the 695 affected genes in the total of biclusters). Meanwhile, the seed nodes are colored on a scale of red, with darker shades indicating their presence in a higher number of biclusters. Node size is determined by eigenvector centrality. Labels are displayed only for seed nodes that appear in 3 or more biclusters and for any node with an eigenvector centrality equal to or greater than 0.3.

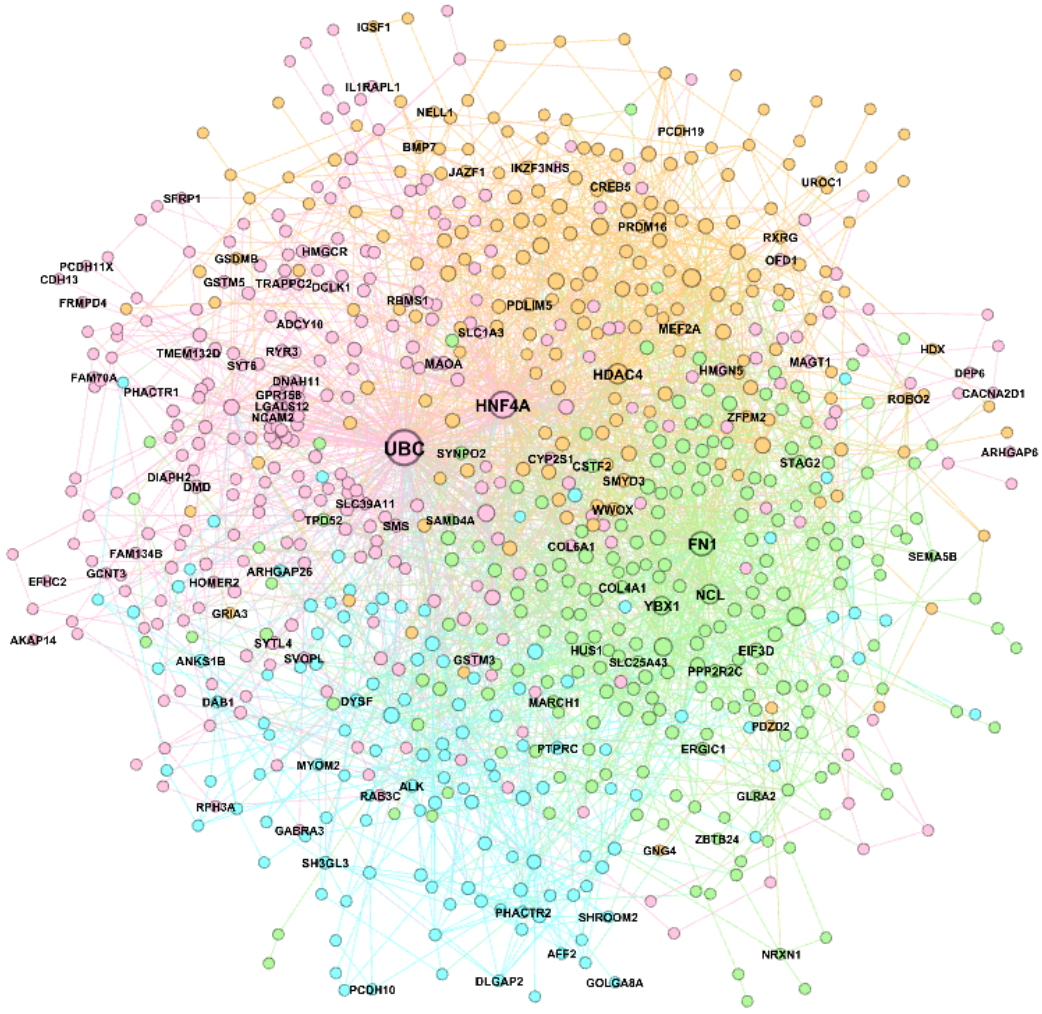
The comprehensive **topological and centrality analysis of the network** can be found in Appendix C.2.1. This analysis emphasizes the **central importance of the UBC gene** in the network despite not being a seed node. It is the node with the highest degree, eigenvector, and betweenness centralities. Specifically, it has a degree of 232, while the average degree is just

6.184, identifying it as a hub. Moreover, the Cytoscape circular layout allowed us to detect that UBC is **connected to 28 seed nodes** that have no other neighbors (Appendix C.2.1, Figure 45).

According to GeneCards, the UBC gene encodes a **polyubiquitin precursor**. Ubiquitin is a small protein that carries out its functions through conjugation with a wide range of target proteins. Conjugation of ubiquitin monomers or polymers can lead to various effects within a cell, such as protein degradation, DNA repair, cell cycle regulation or regulation of other cell signaling pathways.

Additionally, the web server of **NetworkAnalyst** itself allows conducting an **ORA** of the genes in the graph, by choosing the desired database or ontology. A brief selection of the most over-represented terms is provided in Appendix C.2.2 (Table 17), along with explanations of their biological significance. However, the annotation work concludes that these terms seem to be associated with cancer biology in general.

The **Louvain algorithm for community detection** provided insights into the modular structure of the network. Using a resolution of 1.7, a network partition into **4 communities** was created. The **modularity** of this partition is **0.388**. In practice, it has been observed that a value higher than 0.3 is a good indicator of modular structure in the graph [73]. Figure 21 displays the same network visualization as Figure 20, but in this case, node colors indicate **membership** in one of the 4 communities. The legend also shows the percentage of nodes in each community.



((a))

3	(36,71 %)
2	(25,76 %)
1	(22,82 %)
4	(14,71 %)

((b))

Figure 21. Minimum protein-protein interaction network derived from the total set of affected genes. Each node color indicates membership in a different Louvain community. Interactions were obtained using NetworkAnalyst, and visualization was done with Gephi. Node size is determined by eigenvector centrality. Labels are displayed only for seed nodes that appear in 3 or more biclusters and for any node with an eigenvector centrality equal to or greater than 0.3.

Cytoscape enabled the identification of **over-represented terms within each community separately**. A manual selection was performed, focusing on non-redundant terms with lower adjusted p-values. The chosen terms are presented in Table 8. Each community exhibits different

over-represented terms, suggesting that many protein-protein interactions occurring within the same community enable the execution of a specific biological function.

Table 8. Selection of over-represented terms (according to Cytoscape) in each Louvain community of genes. The p-value for each term is displayed next to it in parentheses.

Community	**Over-represented terms**
1	Transcription by RNA polymerase II (3.58E-58)
2	Extracellular exosome [27] (5.87E-15)
3	Synapse (4.02E-12) - Cell junction [28] (2.43E-11)
4	Cell junction (1.20E-26) - ErbB signaling pathway [29] (1.59E-19)

The **Circle Pack topology** of Gephi enabled the visualization of nodes grouped by communities (Figure 22). Indicating seed nodes in reddish tones, as in the original visualization, we identified community 3 as the community with the highest proportion of seed nodes. This is likely influenced by the fact that it includes the UBC gene. Community 3 is precisely the one showing over-representation of the term "synapse" (Table 8), which aligns with the results of the custom-designed ORA conducted on the total of 695 affected genes and on those directly related to CRC.

[27] Exosomes are extracellular vesicles transporting different molecules between cells [74].

[28] Cell junctions are multiprotein complexes facilitating adhesion between neighboring cells or between a cell and the extracellular matrix [75].

[29] The ERBB2 signaling pathway is a crucial signaling pathway playing a major role in various cellular responses such as proliferation and differentiation [76].

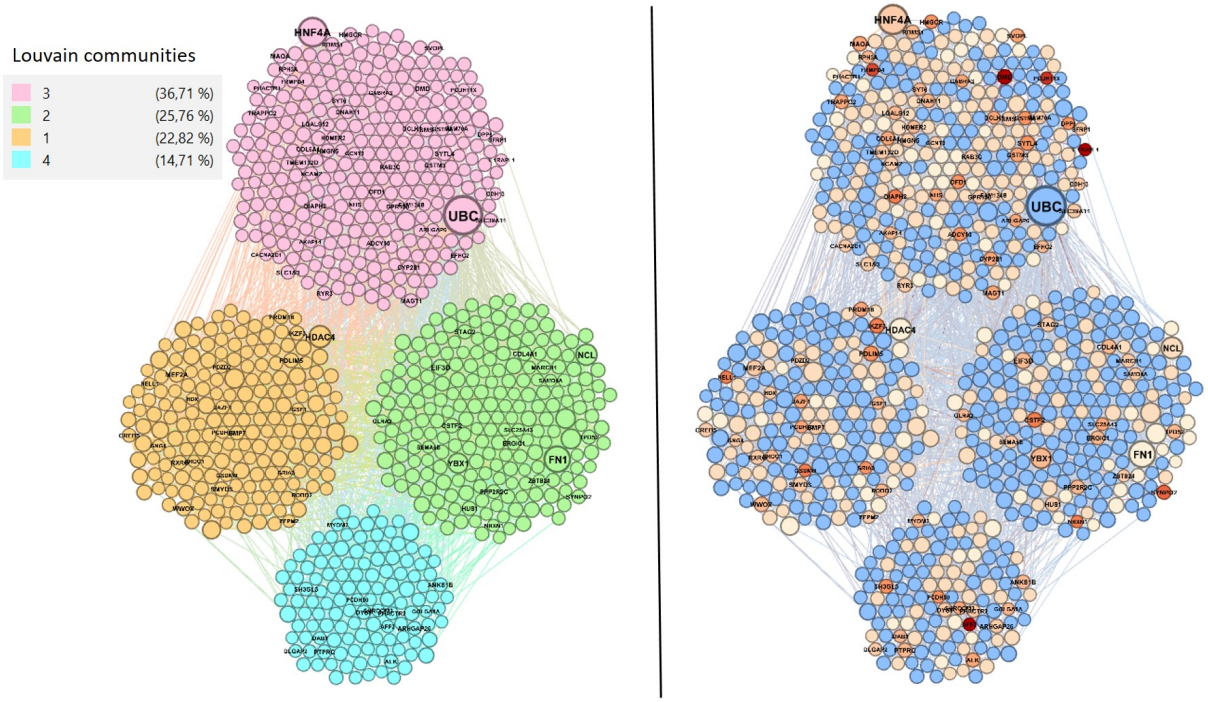


Figure 22. Minimum protein-protein interaction network derived from the total set of affected genes: Circle Pack layout based on community membership. Interactions were obtained using NetworkAnalyst, and visualization was done with Gephi. In the left image, each node color indicates membership in a different Louvain community. In the right image, the blue nodes represent non-seed nodes, while the seed nodes are colored on a scale of red, with darker shades indicating their presence in a higher number of biclusters. Node size is determined by eigenvector centrality. Labels are displayed only for seed nodes that appear in 3 or more biclusters and for any node with an eigenvector centrality equal to or greater than 0.3.

3.4.3 Intergenic SNPs Analysis.

Intergenic SNP Ensembl Table.

Most Ensembl attributes considered key in genic SNP annotation (those shown in Figure 15) are gene-specific and, therefore, useless for intergenic SNP analysis. Consequently, Figure 23 only displays the chromosomal distribution of intergenic SNPs. This chart reveals that, similar to what was observed for genic SNPs, **chromosome X** accumulates the most intergenic SNPs and **Chromosomes 9 and Y** have the fewest SNPs of this type. However, in this case, the second most frequent chromosome is **chromosome 6**, not chromosome 1.

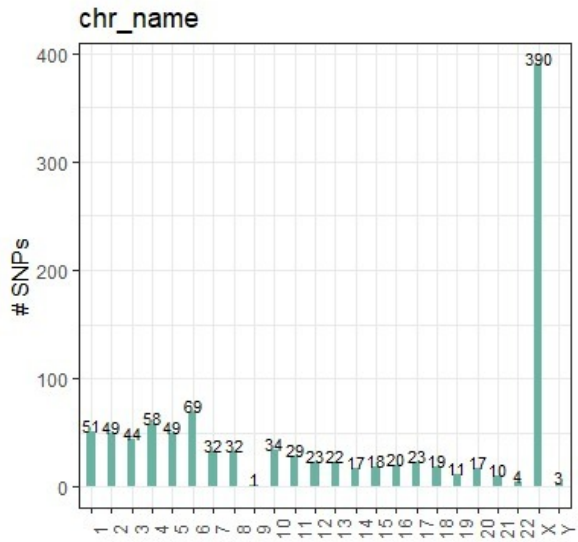


Figure 23. Bar chart depicting the Ensembl "chr_name" attribute in the intergenic SNP Ensembl table.

Despite the lack of information regarding **intergenic SNPs** in Ensembl, **RegulomeDB** facilitated the prioritization of intergenic SNPs that may have a greater implication in the disease. **47 SNPs passed the filter** with a probability of **0.25** or higher of having a specific regulatory function in the colon, intestine, or large intestine. Table 9 displays their identifiers, the overall probability of being regulatory, their ranking score in the evidence of their regulatory function, the chromosome they are located on, and the three regulatory tissue-specific probabilities.

Table 9. Intergenic SNPs with a probability of 0.25 or higher of having regulatory function in the colon, large intestine or intestine according to RegulomeDB.

SNP	reg. prob	ranking	chrom	colon	intestine	large intestine
rs503832	0.969	2b	chr1	0.287	0.287	0.287
rs2482806	0.949	5	chr1	0.28	0.28	0.28
rs12023396	1	2a	chr1	0.296	0.296	0.296
rs1902719	1	3a	chr10	0.296	0.296	0.296
rs6584977	0.974	3a	chr10	0.288	0.288	0.288
rs6578958	0.857	2b	chr11	0.253	0.253	0.253
rs12420118	1	5	chr11	0.296	0.296	0.296
rs4405339	1	5	chr11	0.296	0.296	0.296
rs7933981	0.86	3a	chr11	0.254	0.254	0.254
rs7295116	0.97	5	chr12	0.287	0.287	0.287
rs11068091	0.894	2b	chr12	0.264	0.264	0.264
rs7328572	0.922	5	chr13	0.273	0.273	0.273
rs7996239	0.874	3a	chr13	0.258	0.258	0.258
rs11624310	1	5	chr14	0.296	0.296	0.296
rs8008466	0.927	1b	chr14	0.274	0.274	0.274
rs12946510	0.998	1b	chr17	0.295	0.295	0.295
rs7232298	0.99	5	chr18	0.293	0.293	0.293
rs12327886	1	2b	chr19	0.296	0.296	0.296
rs10414428	0.861	3a	chr19	0.254	0.254	0.254
rs10182660	0.922	5	chr2	0.273	0.273	0.273
rs12158877	0.907	2c	chr22	0.268	0.268	0.268
rs6782157	0.99	1a	chr3	0.293	0.293	0.293
rs115304889	1	5	chr3	0.296	0.296	0.296
rs1566485	1	2a	chr4	0.296	0.296	0.296
rs11730206	0.968	2a	chr4	0.286	0.286	0.286
rs4518223	0.866	2b	chr4	0.256	0.256	0.256
rs258494	0.985	5	chr5	0.291	0.291	0.291
rs17454570	1	3b	chr5	0.296	0.296	0.296
rs13156607	0.873	3a	chr5	0.258	0.258	0.258
rs9466936	0.9	2c	chr6	0.266	0.266	0.266
rs2516491	0.851	3a	chr6	0.251	0.251	0.251
rs9472158	1	2b	chr6	0.296	0.296	0.296
rs9386492	0.922	5	chr6	0.273	0.273	0.273
rs7793526	0.847	3a	chr7	0.25	0.25	0.25
rs10239366	0.93	2a	chr7	0.275	0.275	0.275
rs62502605	0.985	3a	chr7	0.291	0.291	0.291
rs12542934	0.849	3a	chr8	0.251	0.251	0.251
rs973815	0.931	6	chr8	0.275	0.275	0.275
rs10102728	0.861	3a	chr8	0.254	0.254	0.254
rs4825757	0.847	3b	chrX	0.25	0.25	0.25
rs6640653	0.851	3a	chrX	0.251	0.251	0.251
rs5948742	0.94	5	chrX	0.278	0.278	0.278
rs697496	0.982	5	chrX	0.29	0.29	0.29
rs5950432	0.874	3a	chrX	0.258	0.258	0.258
rs7050010	0.941	2b	chrX	0.278	0.278	0.278
rs5922152	0.851	3a	chrX	0.251	0.251	0.251
rs5930119	0.979	5	chrX	0.289	0.289	0.289

In this table, it can be quickly observed that none of the intergenic SNPs exhibit a very high probability of having a specific regulatory function in the organs of interest. Instead, it seems that having such a **high probability** of being **regulatory in general** (as shown in the first column of the table) increases the probability of having some regulatory function specific to the colon, but possibly also increases the probability of having a specific function in **any other organ**.

Global Analysis Summary: taking everything into account, the analysis of the complete sets of SNPs and CRC patients has facilitated the examination of general characteristics across the entire sample. The results pointed towards the genetic origins of medical conditions (neurological disorders and hypercalcemia) already known to be prevalent in CRC patients.

3.5 Bicluster-Specific Analyses.

As mentioned earlier, the majority of the data analysis conducted with the entire set of patients and SNPs was replicated individually for each of the 16 biclusters. Furthermore, distinct information emerged from these bicluster-specific analyses, including gene maps and genotypic frequencies of SNPs. Consequently, the resulting information is extensive and cannot be fully accommodated in this thesis document. However, it can be found in the HTML document accessible through this URL: `https://github.com/Almorox/Masters_Thesis_Characterization_Cancer_Groups/blob/main/Global_and_Bicluster_Specific_Analyses_CRC_Results.zip`

3.5.1 Biclusters Overview: Patient, SNP, and Gene Data.

Although we will not include or discuss the analysis of each bicluster, Figure 24 summarizes the **general characteristics** (expressible in numerical format) extracted from each one. These characteristics can be related to **patients** (clinical information), **SNPs**, or **genes** affected by the SNPs. Furthermore, **bicluster-specific** values can be compared with those of the total set of SNPs or patients (**bicluster-no-specific**), which are shown in the first row of the table.

General and bicluster-specific information

	Patients					SNPs			Affected Genes						
Bicluster	# Patients	% Metast.	Mean age	% Females	% Surg.	# SNPs	% Interg.	% Genic	# Genes	% CRC dir.	% CRC ind.	% C. dir.	% C. ind.	% No rel.	% DEGs
All patients	73	34.72	63.38	21.92	72.60	1997	51.33	48.67	695	21.01	59.57	3.60	3.45	12.37	10.36
2.2	35	29.41	61.73	8.57	68.57	1471	49.29	50.71	546	21.43	58.79	2.93	4.03	12.82	11.17
3.2	3	66.67	63.67	0.00	66.67	1439	49.97	50.03	532	21.24	61.47	2.63	3.20	11.47	11.47
8.4	54	37.74	63.13	12.96	72.22	181	54.70	45.30	58	20.69	56.90	5.17	1.72	15.52	6.90
10.1	34	29.41	64.26	11.76	76.47	171	49.71	50.29	55	20.00	56.36	7.27	3.64	12.73	12.73
10.4	3	0.00	52.50	33.33	100.00	110	71.82	28.18	16	25.00	56.25	6.25	6.25	6.25	6.25
11.5	35	41.18	63.97	2.86	71.43	341	50.73	49.27	133	16.54	63.16	4.51	3.01	12.78	7.52
12.1	11	54.55	62.82	9.09	81.82	59	52.54	47.46	14	21.43	64.29	7.14	0.00	7.14	7.14
12.9	8	37.50	61.62	0.00	62.50	64	51.56	48.44	19	10.53	63.16	10.53	0.00	15.79	10.53
13.8	27	34.62	64.88	0.00	81.48	154	61.04	38.96	42	11.90	66.67	7.14	4.76	9.52	11.90
13.10	7	42.86	62.71	0.00	71.43	126	50.00	50.00	39	23.08	66.67	2.56	2.56	5.13	17.95
14.4	19	26.32	65.00	0.00	94.74	65	33.85	66.15	17	11.76	76.47	5.88	0.00	5.88	11.76
15.6	11	45.45	64.82	9.09	63.64	229	55.46	44.54	75	21.33	58.67	5.33	1.33	13.33	12.00
16.7	26	40.00	65.04	15.38	65.38	103	57.28	42.72	23	17.39	69.57	4.35	4.35	4.35	17.39
16.12	5	20.00	44.25	0.00	100.00	57	61.40	38.60	16	31.25	62.50	0.00	0.00	6.25	0.00
17.7	12	33.33	65.00	16.67	66.67	172	59.88	40.12	49	14.29	67.35	8.16	4.08	6.12	10.20
17.8	8	62.50	58.14	0.00	75.00	69	82.61	17.39	7	14.29	71.43	14.29	0.00	0.00	14.29

Figure 24. Summary table of each bicluster (patient, SNP, and gene-related information).

Clinical Characteristics:

- The percentage of **metastatic cases** in the total of 73 patients is 34.72%. We can observe that most bicluster-specific values deviating from this value occur in biclusters with few patients, such as 3.2 or 10.4.
- Most biclusters (especially those not having too few patients) have an **average age** close to the bicluster-no-specific value, which is 63.38. The most striking case is **bicluster 16.12**, which has an average age of 44.25.
- Despite the overall **female percentage** being almost 22% in the 73 patients, no bicluster reaches this value.
- Regarding **surgery** percentages, the value in the total number of patients is high and remains so in most biclusters. **Bicluster 14.4** stands out for having 18 out of 19 patients who underwent surgery.

SNP Characteristics:

- We already knew that the number of SNPs included in each bicluster is highly variable, but now we can see that the **percentages of intergenic SNPs** (as opposed to genic SNPs) also vary. For instance, in **bicluster 14.4**, 22.85% of SNPs are intergenic, while in **bicluster 17.8**, this percentage rises to 82.61%.

Gene Characteristics:

- The aforementioned percentage of intergenic SNPs logically causes significant differences in the **number of affected genes** in each bicluster. Continuing with the previous example, biclusters 14.4 and 17.8, despite having a similar number of SNPs, have 17 and 7 affected genes, respectively. It is also important to note that the value of affected genes is influenced by multiple SNPs from the same bicluster being located in a single gene.
- The bicluster with the highest percentage of genes **directly related to CRC** is **16.12**, with 31.25%. On the other hand, if the bicluster-no-specific percentage of genes **indirectly related to CRC** is already high (almost 60%), most biclusters exceed this percentage, reaching its maximum in **bicluster 14.4** (76.47%).
- The percentages of genes directly and indirectly **related to cancer**, as well as the percentages of genes for which **no relationship** with these diseases has been found, are quite variable (as expected, given biclusters with few genes) and generally low.
- Lastly, there seems to be no bicluster with a significant over-representation of genes identified as **DEGs** in our RNA-seq data analysis.

Selection of two interesting biclusters for further analysis.

Based on Figure 24, we believed it would be valuable to **delve deeper** into **biclusters 16.12 and 17.8**. Here are the reasons:

- **Bicluster 17.8** is a bicluster with only **7 affected genes**, implying that PGMRA has grouped 8 patients based on their similarity in SNP genotypes affecting only 7 genes. The fact that these 8 patients share a CRC diagnosis suggests that these genes, and specifically these SNP genotypes, may be linked to the onset and progression of the disease.
- **Bicluster 16.12** is also intriguing for the same reason. Although it has more affected genes, it remains a manageable set for this Master's thesis, with **16 genes** in this case.

- Another characteristic that makes bicluster 16.12 interesting is the **low average age** of its patients, indicating a higher likelihood that genetic factors play a significant role in the development of their disease. It is evident that, over time, any individual (with or without genetic predisposition) will have a higher chance of developing CRC due to the natural aging process and accumulated exposure to environmental factors.
- Additionally, biclusters 16.12 and 17.8 represent patient groups with the **lowest and highest metastasis percentages**, 20% and 62.5%, respectively. Considering the low number of patients in these biclusters, it might seem that these percentages do not deviate from the expectations (being the bicluster-no-specific percentage 34.72%). However, the 3 patients without metastasis in bicluster 17.8 are also present in bicluster 16.12. In other words, **the 5 patients fulfilling "17.8=Yes" and "16.12=No", have metastasis.** All of this occurs in a context where **randomly** selecting **5 patients** from the total of 73 and having all of them with **metastasis** is an event with a probability of **0.0035** ($= \frac{25}{73} \cdot \frac{24}{72} \cdot \frac{23}{71} \cdot \frac{22}{70} \cdot \frac{21}{69}$). This information aligns with the **penultimate association rule** in Table 2, having a lift of 2.92.
- Finally, there is relatively **low similarity** between biclusters 16.12 and 17.8 in terms of **SNP composition**. Only 8.77% of SNPs from 16.12 are also present in 17.8, and 7.25% of SNPs from 17.8 are found in 16.12 (Figure 10). Moreover, they only share one of their genic SNPs, and as a result, they have **only one shared gene**, PCDH11X. This information implies that both biclusters have distinct and small gene networks. They also point towards different sets of intergenic SNPs (which are larger than genic sets, in both biclusters).

3.5.2 Bicluster 16.12.

Bicluster 16.12 Clinical Data.

From the analysis of clinical bar charts in bicluster 16.12 (Appendix C.3.1, Figures 46 and 47), we can extract the following information:

- This bicluster consists of **5 male patients**, among whom **3** are **under 50 years old**.
- Only one of them has **metastasis**, affecting the liver and lymph nodes but not the lungs.
- All of them have undergone **surgery**, and the **tumor locations** vary.
- One is in **stage II**, and the rest are in **stage III** (none, not even the one with metastasis, is in the most advanced stage, stage IV).

- **3** of them have **moderately differentiated tumors**.
- The **basal CEA** levels are low, with only one in the [5-9] interval.

As previously mentioned, the distribution of the "Age_Range" attribute is quite unexpected, since 3 of the 6 total patients under 50 years old are included in bicluster 16.12. However, other attribute-value pairs that appear frequently in the bicluster, such as "Stage=III" or "degree=moderately differentiated", also appear frequently in bicluster-no-specific clinical data.

Figure 25 emphasizes the difference in the **"Age_Range"** attribute distribution between patients belonging to bicluster 16.12 and those who do not. Additionally, we observed that the only patient over 50 in the bicluster is **54 years old**, remaining one of the youngest in the age range [50-70). Consequently, bicluster's mean age remains low (44.25) compared to that of the total set of patients (63.36).

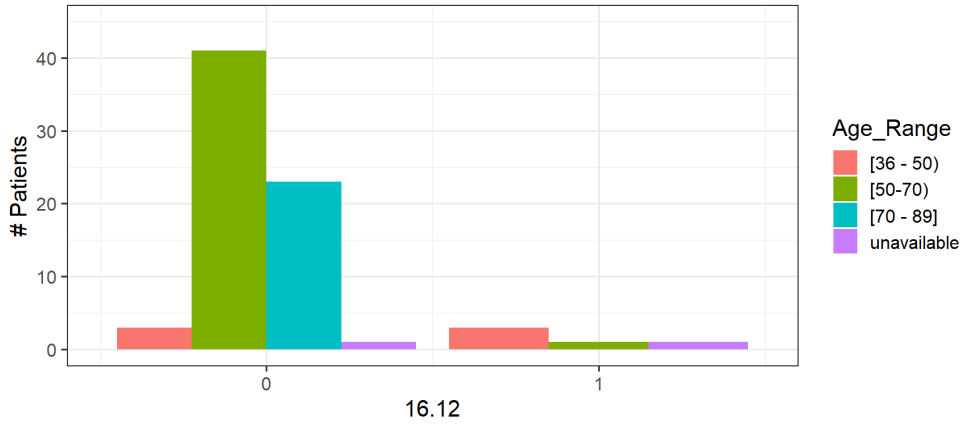


Figure 25. Distribution of the variable "Age_Range" based on membership or non-membership to bicluster 16.12.

As mentioned in the introduction, **inflammatory bowel disease (IBD)** is a risk factor for developing CRC. IBD is a term for two conditions (Crohn's disease and ulcerative colitis) that are characterized by chronic inflammation of the gastrointestinal tract [77]. Patients with extensive IBD, particularly patients with ulcerative colitis, have an up to **20-fold increased risk** of developing CRC. Furthermore, the age of onset is thought to be approximately **20 years younger** than sporadic CRC in the general population [78]. This information leads us to consider the possibility that the 5 patients belonging to bicluster 16.12 are cases of **IBD-associated CRC**, which PGMRA has grouped together based on their genotypic similarity regarding a small set of SNPs. However, we do not have access to such clinical information, so the hypothesis is far

from confirmed.

Figure 26, analogous to the previous one, depicts the difference in the distribution of the **"Metastasic"** attribute between patients belonging to 16.12 and those who do not. We can observe that in bicluster 16.12, the percentage of non-metastatic patients is higher. However, this distribution is **not particularly unexpected**, as the absence of metastasis remains very common among patients excluded from this bicluster.

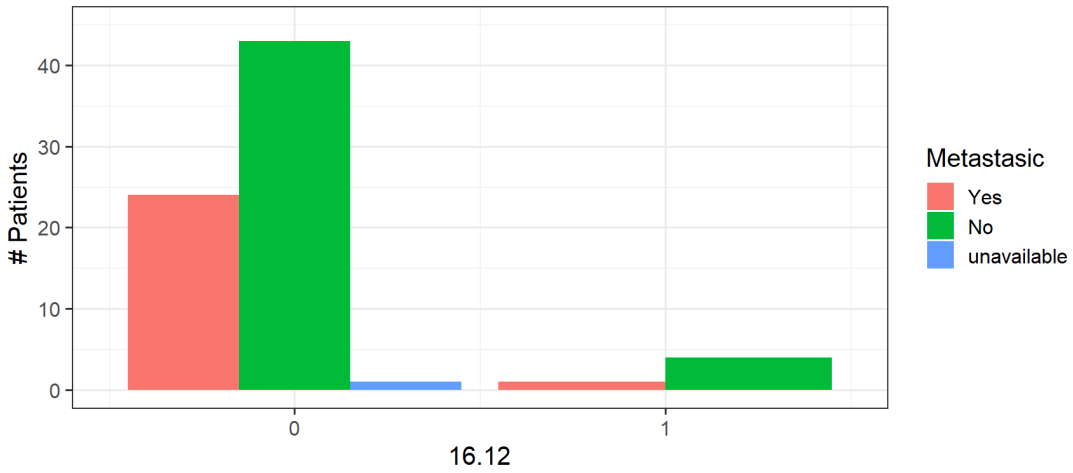


Figure 26. Distribution of the variable "Metastasic" based on membership or non-membership to bicluster 16.12.

Summary Table of Genic SNPs in Bicluster 16.12.

One of the outcomes produced by the function we created to conduct bicluster-specific analyses consists of a **pair of summary tables** outlining general features. One is dedicated to **bicluster genic SNPs**, while the other focuses on **bicluster intergenic SNPs** that were filtered by RegulomeDB.

In the **case of bicluster 16.12**, none of its intergenic SNPs passed the RegulomeDB filter, so only the table of genic SNPs was generated. This table, shown in Figure 27, includes fields related to the affected gene, Ensembl SNP information, and calculations of genotypic percentages in the bicluster. We can observe **22 SNPs** affecting **16 genes**, as 7 SNPs are located in the same gene, IGSF1. Out of the 16 genes, **5 have a direct relation to CRC**, but none of them has been identified as a DEG in our RNA-Seq data analysis. Many of these SNPs are

located on the **X chromosome**, as expected based on the frequency of this chromosome in the bicluster-no-specific analysis (Figure 15). For the same reason, it was also expected that the majority of consequence types of the transcript variant were **intronic**, with the associated **"LOW" impact** value. Regarding the **genotypic percentages**, we can see that there are 10 SNPs for which none of the patients has genotype *1* (major allele homozygote), meaning that all patients in the bicluster have, in at least one dose, the minor allele of the SNP. There are also SNPs for which genotype *3* (recessive minor allele) is the most common. In fact, there are two SNPs that appear with genotype *3* in all of the patients. All this provokes **very high MAFs** ($\frac{\text{chromosomes with the SNP minor allele}}{\text{chromosomes}}$) for many of these SNPs.

Bicluster 16.12 - Genic SNPs											
	Gene info.			SNP - Ensembl info.			SNP - Bicluster 16.12 info.				
	Gene	Rel	DEG	chrom	conseq_type	impact	%_NA	%_g1	%_g2	%_g3	Calc_MAF
rs10152417	RASGRF1	D.CRC	-	15	intron_variant	LOW	0	0	20	80	0.9
rs2859168	CSTF2	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs12009352	PCDH11X	IND.CRC	-	X	intron_variant	LOW	0	0	0	100	1.0
rs6620925	SYTL4	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs10776699	GSTM5	IND.CRC	-	1	intron_variant	LOW	0	0	20	80	0.9
rs1332018	GSTM3	D.CRC	-	1	5_prime_UTR_variant	LOW	0	0	20	80	0.9
rs12305014	PDE6H	IND.CRC	-	12	intron_variant	LOW	0	0	80	20	0.6
rs2475410	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs2503357	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs2503359	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4240130	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs4415478	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs5930459	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs5932901	IGSF1	IND.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4458170	ALK	D.CRC	-	2	intron_variant	LOW	0	0	60	40	0.7
rs222364	MTM1	D.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4825330	NHS	IND.CRC	-	X	intron_variant	LOW	0	20	0	80	0.8
rs1902957	DMD	D.CRC	-	X	intron_variant	LOW	0	40	0	60	0.6
rs4828954	RP11-40F8.2	IND.CRC	-	X	intron_variant	LOW	0	0	0	100	1.0
rs4970777	RP4-735C1.4	IND.CRC	-	1	intron_variant	LOW	0	0	20	80	0.9
rs13108021	RP11-395F4.1	NO REL.	-	4	intron_variant	LOW	0	0	40	60	0.8
rs10454254	F11-AS1	IND.CRC	-	4	intron_variant	LOW	0	0	80	20	0.6

Figure 27. Summary table of genic SNPs in bicluster 16.12. "Rel" represents gene relationships (direct or indirect) with CRC or cancer. "conseq_type" represents the SNP Ensemble attribute "consequence_type_tv". "%_NA" represents the percentage of patients with missing genotypes. "%_g1" represents the percentage of patients with genotype *1*. "Calc_MAF" indicates the Minimum Allele Frequency calculated for the bicluster.

ORA of Affected Genes in Bicluster 16.12.

The function responsible for conducting bicluster-specific analyses also performs an **ORA**, using

the genes affected in the bicluster as input. Figure 28 displays the over-represented **diseases** in the genes affected in bicluster 16.12. There are only three, and at least the first two are directly related to **muscle tissue**: absence of muscle and leiomyosarcoma [30]. The third disease, gigantism, is a disorder caused by excessive **GH (Growth Hormone)** levels [80], leading to excessive growth of bones, muscles, and many internal organs [81].

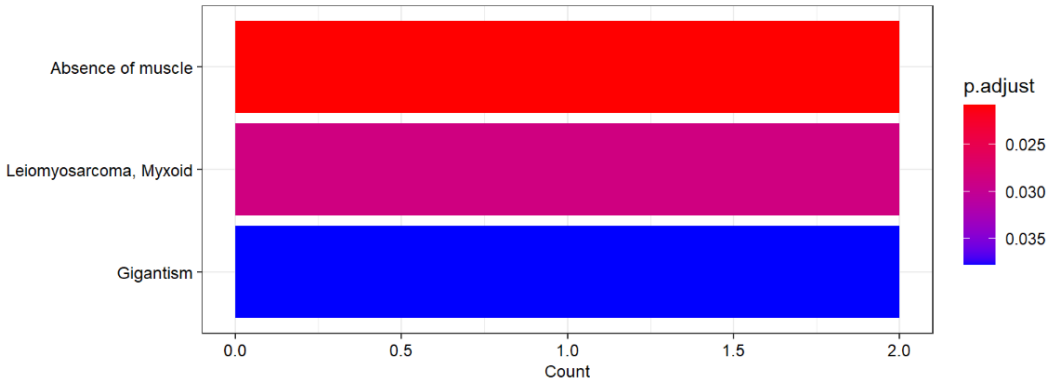


Figure 28. Bar chart of the over-represented diseases in the genes affected in bicluster 16.12. The size of the recognized query gene set in the DisGeNet database was 13.

We have verified that all three diseases are associated with the **DMD** gene. "Absence of muscle" is also linked to **MTM1** , while "gigantism" is also connected to **IGSF1**.

On the other hand, both the **MF** terms in Table 10 and the **pathway** listed in Table 11 are over-represented in bicluster 16.12 genes due to the presence of GSTM3 and GSTM5, encoding two **glutathione S-transferases** of mu class.

Table 10. Over-represented MF GO terms in bicluster 16.12 genes.

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
GO:0004364	glutathione transferase activity	2/11	1.66E-04	8.95E-03	5.20E-03
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	2/11	9.15E-04	4.94E-02	1.45E-02

[30] Leiomyosarcoma is a type of rare cancer that grows in the smooth muscles, which are found in the hollow organs of the body, including the intestines [79]

Table 11. Over-represented pathways in bicluster 16.12 genes.

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
hsa00480	Glutathione metabolism	2/9	1.50E-03	3.60E-02	7.21E-03

The mu class of enzymes functions in the **detoxification** of electrophilic compounds, including carcinogens, by conjugation with glutathione. The genes encoding these enzymes are known to be **highly polymorphic**. These genetic variations can change an individual's susceptibility to carcinogens and toxins. Moreover, tissue-based proteomic studies revealed GSTM3 as a novel marker for regional lymph node metastasis in CRC [82].

We wanted to delve deeper into the function of the 16 genes affected in bicluster 16.12 and found that **several** of them are related to **muscle tissue**, not just the three mentioned earlier. For some, we haven't found a direct connection to muscle, but they are associated with **actin**. Actin is a family of globular, multifunctional proteins that form microfilaments in the **cytoskeleton** [31] and the thin filaments, part of the contractile apparatus in **muscle cells** [84].

The relationships found between these genes and muscle or actin are shown in Table 12. The last column of the table displays the PGMRA biclusters in which the gene is affected. We can observe that most genes are affected in **many other biclusters**, suggesting a potential involvement in CRC for patients beyond those in bicluster 16.12. Only two genes appear exclusively in bicluster 16.12: RASGRF1 and PCDE6H.

[31] The cytoskeleton is a complex, dynamic network of interlinking protein filaments present in the cytoplasm of all cells [83].

Table 12. Genes affected in bicluster 16.12: description, relationship with actin or muscle, additional information, and biclusters in which each gene appears. The source of the data in columns 2-4 is GeneCards, unless otherwise specified in the cell.

Gene	Description	Relationship with muscle or actin.	Additional info	Biclusters
RASGRF1	Ras Protein Specific Guanine Nucleotide Releasing Factor 1	Ras, among other functions, is involved in cytoskeletal actin organization [85].	Promotes the exchange of Ras-bound GDP by GTP.	16.12
CSTF2	Cleavage Stimulation Factor Subunit 2	Related disease: oculopharyngeal muscular distrophy.	Required for polyadenylation and 3'-end cleavage of pre-mRNAs.	2.2, 3.2, 11.5, 13.8, 13.10, 14.4, 16.12
PCDH11X	Protocadherin 11 X-Linked	-	Potential calcium-dependent cell-adhesion protein.	2.2, 3.2, 8.4, 10.1, 10.4, 11.5, 13.8, 15.6, 16.12, 17.8
SYTL4	Synaptotagmin Like 4	-	Involved in intracellular membrane trafficking.	2.2, 3.2, 11.5, 13.8, 14.4, 16.12
GSTM5	Glutathione S-Transferase Mu 5	Regulation of the actin cytoskeleton pathway is strongly impacted by the absence of glutathione [86].	Conjugation of reduced glutathione to hydrophobic electrophiles.	2.2, 3.2, 11.5, 12.1, 14.4, 16.12
GSTM3	Glutathione S-Transferase Mu 3	‖	‖	2.2, 3.2, 11.5, 12.1, 16.12
PDE6H	Phosphodiesterase 6H	-	Amplification of the visual signal.	16.12
IGSF1	Immunoglobulin Superfamily Member 1	Related disease: gigantism (DisGeNet).	Seems to be a coreceptor in inhibin signaling.	2.2, 3.2, 10.4, 16.12

Continued on next page

ALK	ALK Receptor Tyrosine Kinase	Related disease: leiomyosarcoma [87].	Genesis and differentiation of the nervous system.	2.2, 3.2, 11.5, 15.6, 16.12
MTM1	Myotubularin 1	Required for muscle cell differentiation. Related disease: absence of muscle (DisGeNet).	Dual-specificity phosphatase.	10.4, 16.12
NHS	NHS Actin Remodeling Regulator	Involved in cell morphology (integrity of the circumferential actin ring and lamellipod formation).	-	2.2, 10.4, 16.12, 17.7
DMD	Dystrophin	Anchors the extracellular matrix to the cytoskeleton via F-actin. Component of the dystrophin-associated glycoprotein complex which accumulates at the neuromuscular junction (NMJ). Structural function in stabilizing the sarcolemma. Related diseases: gigantism, absence of muscle, leiomyosarcoma. (DisGeNet).	-	2.2, 3.2, 8.4, 10.1, 10.4, 11.5, 12.1, 12.9, 13.8, 15.6, 16.12, 17.7
PTCHD1-AS	PTCHD1 Antisense RNA	-	RNA Gene affiliated with the lncRNA class. In the summary table it appears as RP11-40F8.2	2.2, 3.2, 8.4, 10.1, 10.4, 11.5, 12.1, 13.8, 13.10, 15.6, 16.7, 16.12, 17.7

Continued on next page

RP4-735C1.4		Related to GSTM1.	RNA Gene, affiliated with the lncRNA class. Genecards lists it as ENSG00000241720.	2.2, 3.2, 11.5, 12.1, 14.4, 16.12
RP11-395F4.1	-	-	RNA Gene, affiliated with the lncRNA class. Genecards lists it as ENSG00000248939	2.2, 3.2, 16.12
F11-AS1	F11 Antisense RNA 1	-	RNA Gene, affiliated with the lncRNA class.	2.2, 3.2, 16.12

Hypotheses Regarding the CRC Molecular Subtype Associated with bicluster 16.12.

We propose that the cases of CRC grouped in bicluster 16.12 may be classified under **Consensus Molecular Subtype 4 (CMS4) - Mesenchymal**, or have a genetic predisposition for it. This subtype is one of the gene expression-based molecular subtypes of CRC in the classification by Guinney *et al.* (Figure 5).

Epithelial-mesenchymal transition (EMT) refers to a process whereby the adhesive polarity of epithelial cancer cells dissipates and changes to mesenchymal cells [88]. This is associated with cadherins [32] loss, breakdown of cell-cell junctions and acquisition of a polarized actin cytoskeleton assembly into invasive pseudopodial protusions [33]. Mesenchymal transition enables invasive cells to navigate through the extracellular matrix and into the vasculature [90]. Figure 5 also indicates that other characteristics specific to this subtype include the activation of TGF-B (Transforming Growth Factor Beta) and the overexpression of genes regulating inflammation.

The **evidence** linking the genes affected in **bicluster 16.12** to molecular functions altered in **CMS4** includes the following:

- **PCDH11X** encodes a protocadherin, with potential functions in cell-cell junctions and the establishment of polarity in the epithelium.
- Several affected genes in the bicluster are associated with actin, potentially causing changes

[32] Cadherins are transmembrane proteins mediating cell-cell adhesion and maintaining epithelial morphology. [89]

[33] A pseudopodial protrusion is a temporary arm-like projection of a eukaryotic cell membrane that emerges in the direction of movement. This phenomenon has long been associated with tumor cell migration and invasion. [90]

in cytoskeletal organization. Notably, **NHS** (involved in lamellipod formation) and **DMD** (interacts with the extracellular matrix), both of which could promote invasion.

- **GSTM3** and **GSTM5** are genes that influence the inflammatory response.
- **DMD** and **MTM1** are related to gigantism, a condition caused by high levels of GH. In the kidneys, GH has been shown to induce TGF-B [91].

Additionally, genes related to actin or muscle tissue may, in turn, be associated with the **high prevalence of sarcopenia** (muscle function loss) in CRC patients (which ranges between 12% and 60% [92]).

On the other hand, the hypothesis that patients in bicluster 16.12 are cases of **IBD-associated CRC** is not ruled out, as multiple studies on IBD have clearly demonstrated **EMT** occurring in IBD [93]. In fact, the prevalence of **sarcopenia** in patients with IBD is also very high (52% in Crohn's disease and 37% in ulcerative colitis [94]). Moreover, preliminary canonical pathway analyses identified the **actin-cytoskeleton pathway** as the most relevant and significantly disrupted in the progression from normal colonic mucosa, to IBD-associated CRC [78].

Protein-Protein Interaction Network of Bicluster 16.12.

Finally, we wanted to explore whether the protein-coding genes in bicluster 16.12 (the first 12 in Table 12) are part of the same general protein-protein interaction network. To do this, we again used NetworkAnalyst. In this case, we created a degree-1 network, but not a minimum one. This means that all the neighbors of each seed node are included, even if these neighbors are not connected to any other seed node. Figure 29 shows the obtained network, with seed nodes in green and non-seed ones in purple. Also highlighted are the seed genes for which we have found a direct relationship with CRC.

We can observe that with a degree-1 network, **9 out of the 12 protein-coding genes** in the bicluster are included in the same connected component. The close **physical proximity** of the gene products implies a heightened likelihood that the SNPs within this bicluster may collaborate, increasing the probability of generating a cumulative effect. However, the most interesting protein-protein interactions in this regard would be those in which UBC is not involved. As we have already seen, UBC interacts with many proteins that do not interact and are unlikely to collaborate with each other. It is also important to remember that for genic SNPs to have

any effect, the presence of **intergenic SNPs** from this same bicluster could be crucial, and unfortunately, we have not been able to analyze them in as much depth.

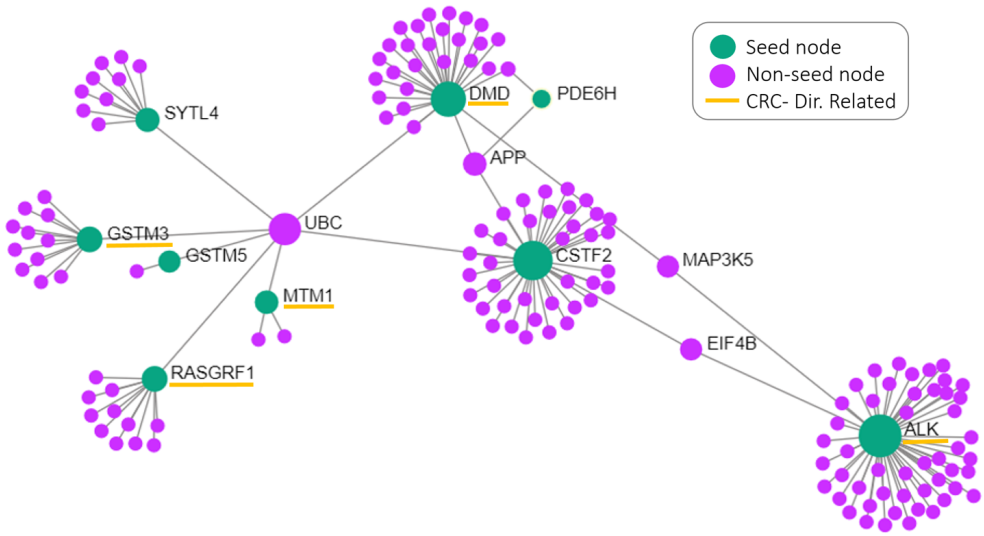


Figure 29. Degree 1 general protein-protein interaction network created from the genes affected in bicluster 16.12. The green nodes represent seed nodes (i.e., genes affected in bicluster 16.12). All nodes within one link's distance of seed nodes are included, colored in purple. The seed nodes corresponding to genes directly related to CRC are underlined in yellow.

3.5.3 Bicluster 17.8.

Bicluster 17.8 Clinical Data.

The distribution of clinical attributes in bicluster 17.8 (Appendix C.3.2, Figures 48 and 49) informs us that:

- This patient group is composed **8 males** distributed across the **three age ranges**. Three of them are aged 70 or older.
- **5** of them have **metastasis**, affecting the liver in 4 patients, the lungs in one, and the lymph nodes in 2.
- The majority have undergone **surgery**.
- **6** of them have **"Loc_T2=rectum-sigmoid"**.
- Patients from bicluster 16.12 were in stage II or III, while those from 17.8 are in **stage III or IV**.
- The most common **degree** is "moderately differentiated", although there are 3 missing values for this variable.
- Two patients have a **basal CEA** of 10 or higher.

Comparing the clinical profiles of **16.12 and 17.8**, it can be interpreted that the latter describes a more **advanced state** of CRC. Especially for the 5 patients in 17.8 who do not belong to 16.12:

- The only patients in bicluster 17.8 without metastasis or aged under 50 are shared patients with bicluster 16.12.
- On the other hand, patients from bicluster 17.8 not included in bicluster 16.12, all have metastasis, and clinical attributes that were not observed in 16.12: stage IV or high CEA values.

The **prevalence of metastasis in each of these bicluster could be a consequence of the age** of their members rather than of their genetic profiles.

Figures 30 and 31 illustrate the distribution of the **"Age_Range"** and **"Metastasic"** variables, respectively, depending on membership in bicluster 17.8. In Figure 30, it is evident that, among patients within the bicluster, the [70-89] age range is more prevalent than the [50-70) range, unlike patients outside this bicluster. Turning to Figure 31, we note that metastasis is more prevalent than the absence of metastasis, a pattern not observed in patients who do not belong to bicluster 17.8. Both images depict significant contrast compared to what we observed for bicluster 16.12 (figures 25 and 26), even though both groups share three patients.

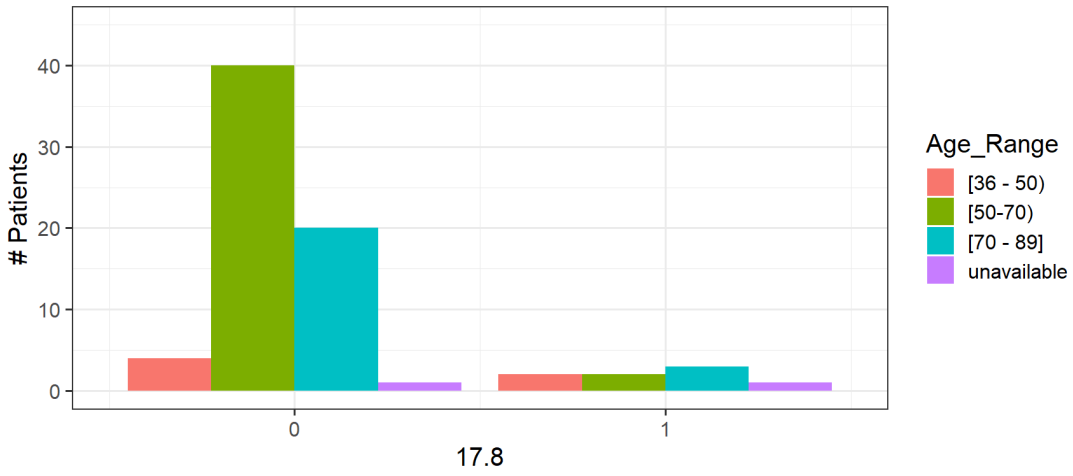


Figure 30. Distribution of the variable "Age_Range" based on membership or non-membership to bicluster 17.8.

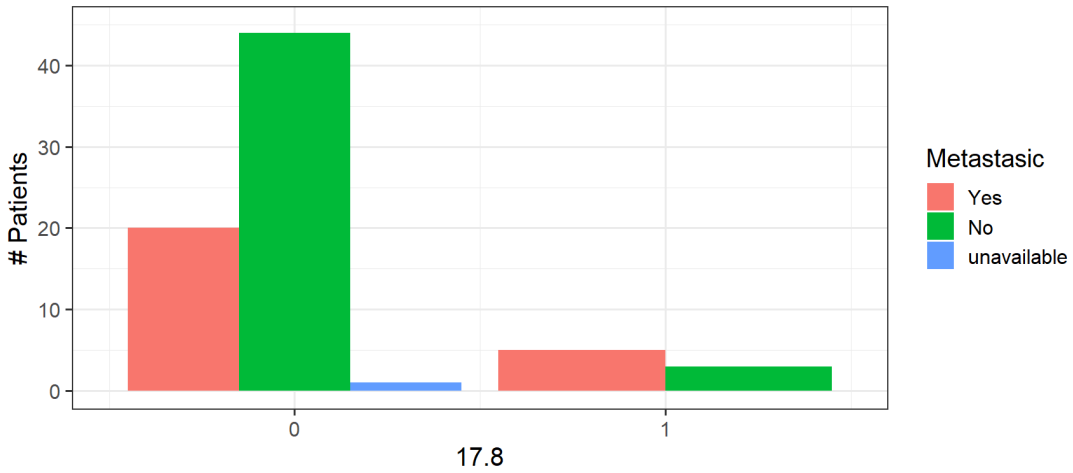


Figure 31. Distribution of the variable "Metastasic" based on membership or non-membership to bicluster 17.8.

Summary Tables of Genic and Intergenic SNPs in Bicluster 17.8.

Among the **intergenic SNPs** in bicluster 17.8, **one was filtered** for having a probability of being regulatory in the colon (or "intestine" or "large intestine") equal to or greater than 0.25, according to RegulomeDB. This SNP is **rs5930119**, and as shown in Figure 32, it has a 98% probability of having a general regulatory function, independently of tissue specificity. The figure also informs us that 5 out of 8 patients (62.5%) in the bicluster have genotype *3* for this SNP.

By consulting the **RegulomeDB v.2 web server**, we examined experiments providing evidence for the regulatory function of the SNP. A positive result in **Chip-seq** indicates the binding of the **NFE2 protein** to the DNA region where this SNP is located. This protein plays a role in red blood cells and megakaryocytic [34] differentiation [95]. Additionally, it has been linked to the regulation of the basal expression of **GSS (glutathione synthetase)** by binding to two motifs in the promoter of this enzyme gene [96]. This could imply another relationship with glutathione metabolism [35] even though there are no genes apparently related to it in this bicluster.

[34] A megakaryocytic is a platelets precursor cell.

[35] For example, if the region where the SNP is located retains the NFE2 protein, preventing it from affecting the expression of the glutathione synthetase.

Bicluster 17.8 - Filtered intergenic SNPs											
	RegulomeDB info.						Bicluster 17.8 info.				
	reg_prob	ranking	chrom	colon	intest.	large_intest.	%_NA	%_g1	%_g2	%_g3	Calc_MAF
rs5930119	0.98	5	chrX	0.29	0.29	0.29	0	37.5	0	62.5	0.62

Figure 32. Summary table of filtered intergenic SNPs in bicluster 17.8. "%_NA" represents the percentage of patients with missing genotypes. "%_g1" represents the percentage of patients with genotype *1*. "Calc_MAF" indicates the Minimum Allele Frequency calculated for the bicluster.

Figure 33 illustrates the summary for the **12 genic SNPs** in bicluster 17.8. We can observe that these affect only **7 genes**, as 5 SNPs are located within the GRIA3 gene, and 2 within the RP5-964N17.1 gene. Only one of the genes, MAGT1, is **directly linked to CRC**, and just one, GRIA3, has been identified as a **DEG** (specifically, as a downregulated gene) in our RNA-seq data analysis. As observed for bicluster 16.12, in this case, we also note a high frequency of the **X chromosome**. At the transcript level, the most severe consequence that all these SNPs can produce is an **intronic variant**.

Concerning **genotypes**, overall, the prevalence of the minor alleles in the bicluster is high. However, differences are observed among various SNPs. For instance, the two SNPs in the RP5-964N17.1 gene have a MAF of 0.5, while the SNP rs5911609 appears in all patients with genotype *3*, consequently exhibiting a MAF of 1.

Bicluster 17.8 - Genic SNPs											
	Gene info.			SNP - Ensembl info.			SNP - Bicluster 17.8 info.				
	Gene	Rel	DEG	chrom	conseq_type	impact	%_NA	%_g1	%_g2	%_g3	Calc_MAF
rs35658443	MAGT1	D.CRC	-	X	intron_variant	LOW	0	25.0	0.0	75.0	0.75
rs12009352	PCDH11X	IND.CRC	-	X	intron_variant	LOW	0	12.5	0.0	87.5	0.88
rs1800654	GRIA3	IND.CRC	DOWN	X	intron_variant	LOW	0	12.5	0.0	87.5	0.88
rs2227098	GRIA3	IND.CRC	DOWN	X	intron_variant	LOW	0	12.5	0.0	87.5	0.88
rs4825857	GRIA3	IND.CRC	DOWN	X	intron_variant	LOW	0	12.5	0.0	87.5	0.88
rs5911598	GRIA3	IND.CRC	DOWN	X	intron_variant	LOW	0	12.5	0.0	87.5	0.88
rs5911609	GRIA3	IND.CRC	DOWN	X	intron_variant	LOW	0	0.0	0.0	100.0	1.00
rs40634	CACNA2D1	IND.CRC	-	7	intron_variant	LOW	0	12.5	25.0	62.5	0.75
rs1538296	SMYD3	IND.CRC	-	1	intron_variant	LOW	0	0.0	37.5	62.5	0.81
rs5929349	RP5-964N17.1	IND.CRC	-	X	intron_variant	LOW	0	50.0	0.0	50.0	0.50
rs6643336	RP5-964N17.1	IND.CRC	-	X	intron_variant	LOW	0	50.0	0.0	50.0	0.50
rs4271959	PCAT4	D.C.	-	4	intron_variant	LOW	0	12.5	37.5	50.0	0.69

Figure 33. Summary table of genic SNPs in bicluster 17.8. "Rel" represents gene relationships (direct or indirect) with CRC or cancer. "conseq_type" represents the SNP Ensemble attribute "consequence_type_tv". "%_NA" represents the percentage of patients with missing genotypes. "%_g1" represents the percentage of patients with genotype *1*. "Calc_MAF" indicates the Minimum Allele Frequency calculated for the bicluster.

ORA of Affected Genes in Bicluster 17.8.

The execution of the ORA, using the 7 genes from bicluster 17.8 as input, yielded over-represented **GO MF** terms exclusively. These terms, associated with trans-membrane transport —specifically, the transport of inorganic cations— are detailed in Table 13. The over-representation stems from 3 out of 5 genes being linked to this activity (2 of the 7 bicluster genes were not recognized in the ontology).

Table 13. Over-represented MF GO terms in bicluster 17.8 genes.

ID	Description	GeneRatio	pvalue	p.adjust	qvalue
GO:0022890	inorganic cation transmembrane transporter activity	3/5	4.15E-04	2.03E-02	2.97E-03
GO:0015318	molecular entity transmembrane transporter activity	3/5	7.06E-04	3.46E-02	2.97E-03

The 3 genes implicated in this activity are **MAGT1**, **GRIA3**, and **CACNA2D1**. Their roles as transporters are detailed in Table 14: they facilitate the transport of **magnesium**, **glutamate**, and **calcium**, respectively. Apart from this specific activity, no other shared functions among genes are immediately apparent .

The **presence** of bicluster 17.8 genes **in other biclusters** is generally lower than that of genes of bicluster 16.12. For instance, CACNA2D1 and SMYD3 appear exclusively in biclusters 2.2 and 3.2, in addition to 17.8. This might suggest that the SNPs associated with these two genes are not linked to the disease as a whole.

Table 14. Genes affected in bicluster 17.8: description, additional information, and biclusters in which each gene appears. The source of the data in columns 2-3 is GeneCards.

Gene	Description	Additional Info	Biclusters
MAGT1	Magnesium Transporter 1	Magnesium cation transporter protein that localizes to the cell membrane. May have a role in N-glycosylation.	3.2, 10.4, 13.10, 16.7, 17.8
PCDH11X	Protocadherin 11 X-Linked	Potential calcium-dependent cell-adhesion protein.	2.2, 3.2, 8.4, 10.1, 10.4, 11.5, 13.8, 15.6, 16.12, 17.8
GRIA3	Glutamate Ionotropic Receptor AMPA Type Subunit 3	Plays an important role in excitatory synaptic transmission.	2.2, 3.2, 10.4, 12.9, 17.7, 17.8
CACNA2D1	Calcium Voltage-Gated Channel Auxiliary Subunit Alpha2delta 1	Mediates the influx of calcium ions into the cell upon membrane polarization. Important role in excitation-contraction coupling.	2.2, 3.2, 17.8
SMYD3	SET And MYND Domain Containing 3	Histone methyltransferase which functions in RNA polymerase II complexes.	2.2, 3.2, 17.8
XACT	X Active Specific Transcript	This gene produces a spliced long non-coding RNA that is thought to play a role in the control of X-chromosome inactivation (XCI).	2.2, 3.2, 10.1, 12.1, 17.8
PCAT4	Prostate Cancer Associated Transcript 4	RNA Gene affiliated with the lncRNA class. Related disease: Prostate Cancer.	3.2, 8.4, 17.8

Protein-Protein Interaction Network of Bicluster 17.8.

Figure 34 depicts the degree-1 protein-protein interaction network derived from NetworkAnalyst using the 5 protein-coding genes of bicluster 17.8 (the first 5 in table 14). The network encompasses **3 of these 5** genes. Notably, **UBC** serves as a shared neighbor between SMYD3 and MAGT1, both additionally connected through **DDOST** (a subunit of the oligosaccharyl transferase complex [97]). Moreover, MAGT1 and CACNA2D1, magnesium and calcium transporters, respectively, are both proteinally linked to **STT3B** (another subunit of the oligosaccharyl transferase complex [98]).

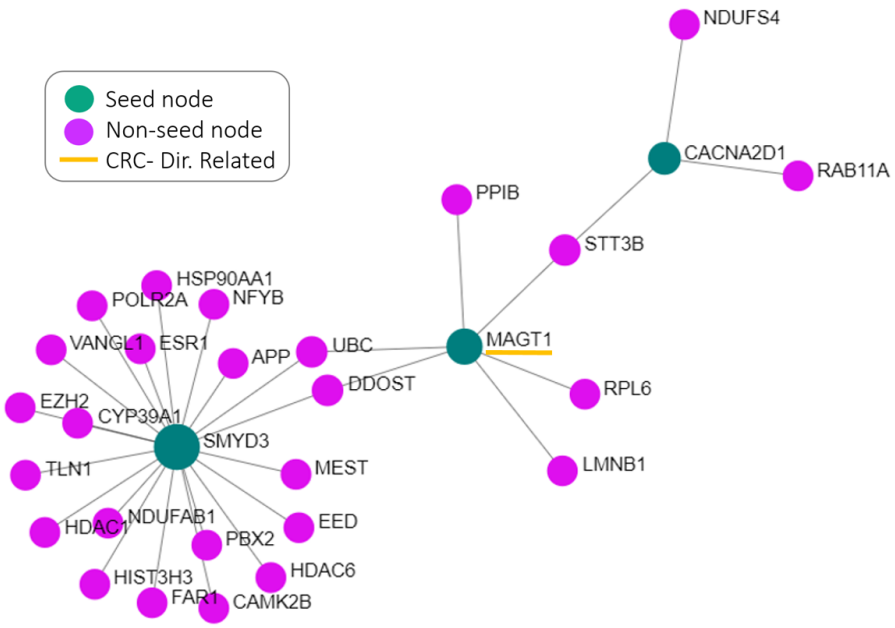


Figure 34. Degree 1 general protein-protein interaction network created from the genes affected in bicluster 17.8. The green nodes represent seed nodes (i.e., genes affected in bicluster 17.8). All nodes within one link's distance of seed nodes are included, colored in purple. The seed nodes corresponding to genes directly related to CRC are underlined in yellow.

In fact, NetworkAnalyst indicated that the term **"oligosaccharyltransferase complex"** from the CC GO (Cellular Component Gene Ontology) is over-represented in this network's gene set. Oligosaccharyltransferase (OST) is a membrane protein complex that carries out **N-glycosylation**.[36]

[36] N-glycosylation is a fundamental post-translational protein modification that is involved in the quality control, trafficking of proteins, signal transduction, and cell-to-cell communication.

3.5.4 General Protein-Protein Interaction Network Created from Genes Affected in either Bicluster 16.12 or 17.8.

To conclude the investigation of these two biclusters, we sought to explore if the affected genes in both of them are interconnected at the protein level. To achieve this, we constructed the minimum network comprising all protein-coding genes from both biclusters, utilizing NetworkAnalyst once again. The outcome is depicted in Figure 35, where genes from bicluster 16.12 are represented in green, those from bicluster 17.8 in blue, and non-seed nodes in purple. **All seed nodes that appeared in the networks specifically created for each bicluster are present**: 9 genes from bicluster 16.12 and 3 from 17.8. Notably, the non-seed gene UBC is connected to all seed nodes except for 2 from 16.12 and 1 from 17.8.

However, there is **no link between seed nodes from different biclusters**. Therefore, these two biclusters do not directly interact at the protein level in any way. Moreover, the gene shared by both, PCDH11X, does not appear in the network.

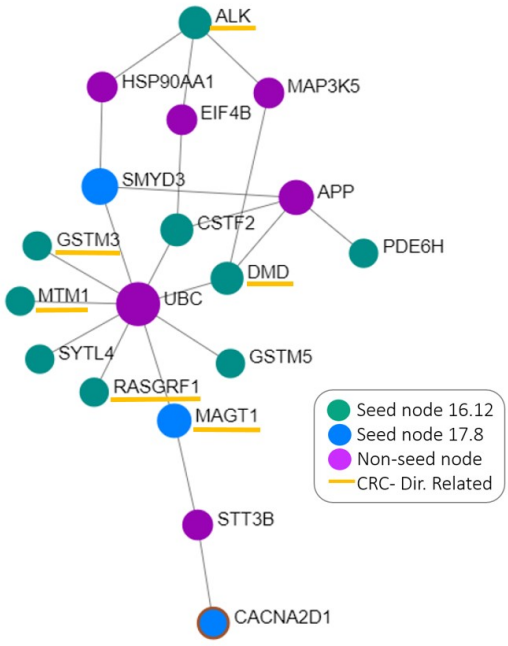


Figure 35. Minimum general protein-protein interaction network of degree 1 created from the genes affected in biclusters 16.12 and 17.8. The green nodes represent seed nodes from bicluster 16.12, while seed nodes from bicluster 17.8 are colored in blue. Additional nodes connecting seed nodes in the graph are included (shown in purple). The seed nodes corresponding to genes directly related to CRC are underlined in yellow.

Bicluster-Specific Analysis Summary: this analysis has enabled the identification of biclusters encompassing few patients and few SNPs. In these, patients exhibit notably high frequencies of the minor alleles, suggesting crucial SNP roles in CRC. Moreover, for bicluster 16.12, potential relationships have been established between clinical characteristics and gene functions, providing new insights into CRC subtypes.

3.6 Identification of Interesting Genes.

From the total set of 695 genes, some appear affected in **many biclusters**, possess **numerous associated SNPs**, or exhibit **high SNP MAFs** in at least one bicluster. Any of these conditions suggests that these genes' SNPs were crucial for grouping CRC patients together (distinguishing them from other diagnoses) and, therefore, could contribute to the disease. Several of these genes' functions were observed to be altered in CRC in other studies or through our own RNA-Seq data analysis, using external samples. In other cases, there is no external evidence of the gene-CRC association, but our results could represent the extraction of novel knowledge.

3.6.1 Genes Affected in Many Biclusters or with Many Associated SNPs.

Table 15 shows the top 5 genes that appear affected in the highest number of biclusters and Table 16 displays the top 5 genes with the highest number of associated SNPs. Of the six different genes that appear in both tables, all are indirectly related to CRC, except for DMD and ADAMTS9-AS1, which are directly related to CRC. Only two of them were identified as DEGs in our RNA-Seq analysis: AFF2 and GRIA3.

Table 15. Top 5 ranking of genes affected in the highest number of biclusters.

Genes	**# biclusters**
PTCHD1-AS	13
AFF2	12
IL1RAPL11	12
DMD	12
ADAMTS9-AS1	11

Table 16. Top 5 ranking of genes affected in the highest number of biclusters

Gene	**# SNPs**
AFF2	16
PTCHD1-AS	14
GRIA3	13
IL1RAPL11	10
DMD	10

The fact that the genes listed in Table 15 are affected in the majority of biclusters suggests that these genes' SNPs may contribute to **common clinical manifestations or symptoms in a substantial portion of our patient sample**. Most of these genes are also present in Table 16 due to their association with numerous SNPs. The only exception is ADAMTS9-AS1, which, interestingly, is associated with only one SNP.

On the other hand, the GRIA3 gene has many associated SNPs (Table 16) but is not affected in as many biclusters —it is involved in 6 biclusters (Table 14), and among them is not 8.4, which includes the highest number of patients. Thus, GRIA3 SNPs may contribute to **specific CRC characteristics in a smaller subset of patients**.

3.6.2 PTCHD1-AS and ADAMTS9-AS1 Genes.

In Table 15 we can observe two genes that are protein-non-coding: both PTCHD1-AS and ADAMTS9-AS1 encode **antisense RNAs (AS-RNAs)**, falling into the category of lncRNAs (long non-coding RNAs). An AS-RNA is a single-stranded RNA that is complementary to a protein-coding mRNA with which it hybridizes, thereby blocking its translation into protein [99].

A study conducted by Chen *et al.* [100] identified **ADAMTS9-AS1 as a prognostic marker in CRC**, promoting cell proliferation and EMT [37]. Our findings complement those of these authors: PGMRA has identified one ADAMTS9-AS SNP, rs9833903, as a relevant variable in CRC patients, pointing towards a **genomic origin of this gene's disruption** in the disease. In several biclusters, the **MAF** of rs9833903 surpasses 0.5. For instance, it is 0.63 in bicluster

[37] ADAMTS9-AS1 targets ADAMTS9 (ADAM Metallopeptidase With Thrombospondin Type 1 Motif 9), which is implicated in the cleavage of proteoglycans, the control of organ shape during development, and the inhibition of angiogenesis [101].

2.2 (with 35 patients) and 0.81 in bicluster 12.9 (with 9 patients). Figure 36 corresponds to the **gene map** of ADAMTS9-AS1, displaying the rs9833903 SNP (this gene map is uniform across all biclusters where the gene is affected).

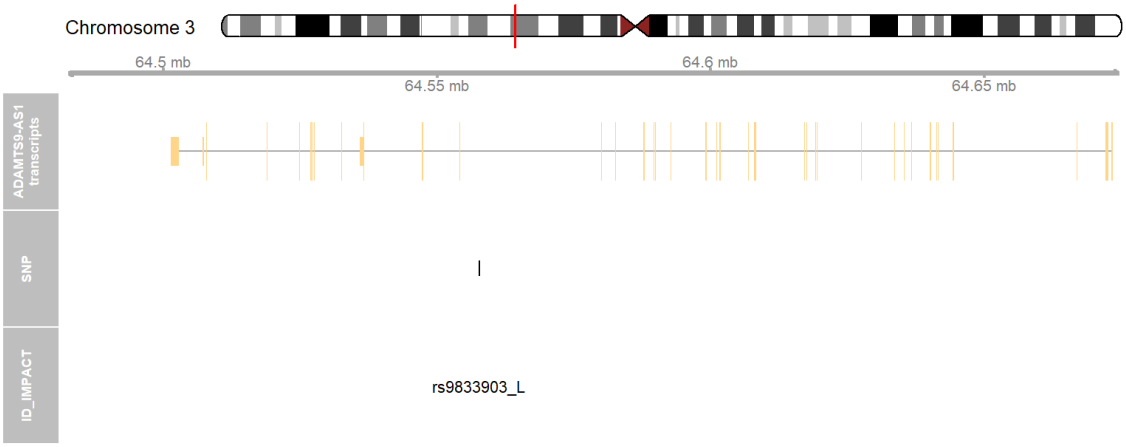


Figure 36. ADAMTS9-AS1 gene map. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

For the **PTCHD1-AS gene**, we have not found external evidence regarding its involvement in the disease. Nevertheless, our results indicate such involvement. The lncRNA encoded by this gene binds to the mRNA encoded by **PTCHD1** [38], which is affected in one of the 16 CRC-biclusters, specifically, in bicluster 3.2. Remarkably, PTCHD1 was identified as a highly **underregulated gene** in CRC, according to our RNA-Seq analysis, with a logFC of -3.66. Therefore, the mechanism of gene expression regulation exerted by PTCHD1-AS on PTCHD1 (illustrated in Figure 37) seems to be highly activated in CRC. This, in turn, could be triggered or influenced by the SNPs' impact in both genes, especially in PTCHD1-AS, as its SNPs appear in the majority of CRC biclusters.

[38] PTCHD1 (Patched Domain Containing 1) encodes a membrane protein. Deletions in this gene are associated with intellectual disability and autism [102].

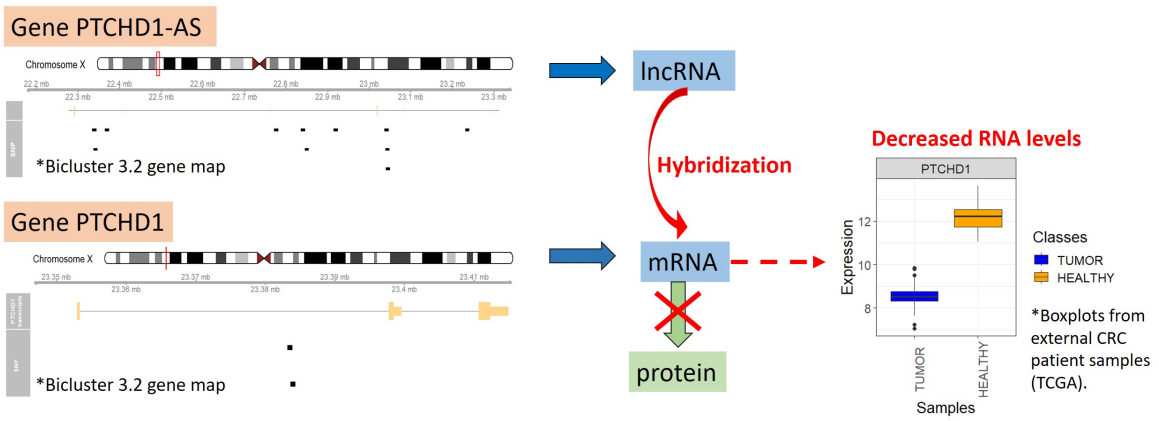


Figure 37. Gene expression regulation exerted by PTCHD1-AS on PTCHD1, which could be influenced in CRC by PTCHD1-AS SNPs.

PTCHD1-AS was one of the genes affected in **bicluster 16.12**, with a single SNP, exhibiting a **MAF of 1** (it appears under the alias RP11-40F8.2 in Table 27). Without reaching this value, the gene also presents **high SNP MAFs in other biclusters**. For example, in bicluster 2.2 (with 35 patients), there are 11 PTCHD1-AS SNPs, as illustrated in Figure 38, and one of them, rs4824190, has a MAF of 0.62.

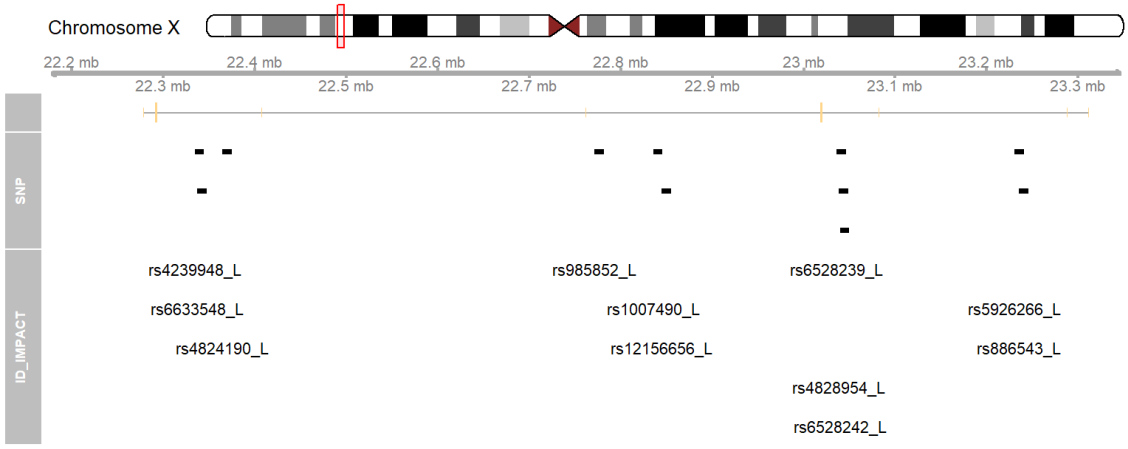


Figure 38. PTCHD1-AS gene map in bicluster 2.2. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

3.6.3 AFF2 Gene.

We also wanted to investigate in more depth the gene with the highest number of associated SNPs, AFF2 (ALF transcription elongation factor 2):

- This gene is affected in **12 biclusters** (2.2, 3.2, 8.4, 10.1, 11.5, 12.9, 13.8, 13.10, 14.4, 15.6, 16.7 and 17.7). Interestingly, it is not affected in any of the two biclusters analyzed in more depth in this study (16.12 and 17.8).
- As we have seen, across all biclusters there are **16 different SNPs** associated with this gene. 14 of them appear simultaneously in bicluster 14.4 (Figure 39).
- It maintains notably **high MAFs in some biclusters**. For instance, in bicluster 14.4 (19 patients), 6 AFF2 SNPs exhibit a MAF of 0.84, as shown in Figure 40.
- It was identified as a **DEG** in our RNA-Seq data analysis, specifically as a downregulated gene (logFC of -2.49).
- According to VarElect, it is **indirectly related to CRC through 5 genes**: PIK3CA, PTEN, IFI27, XRCC2, MIR125A.

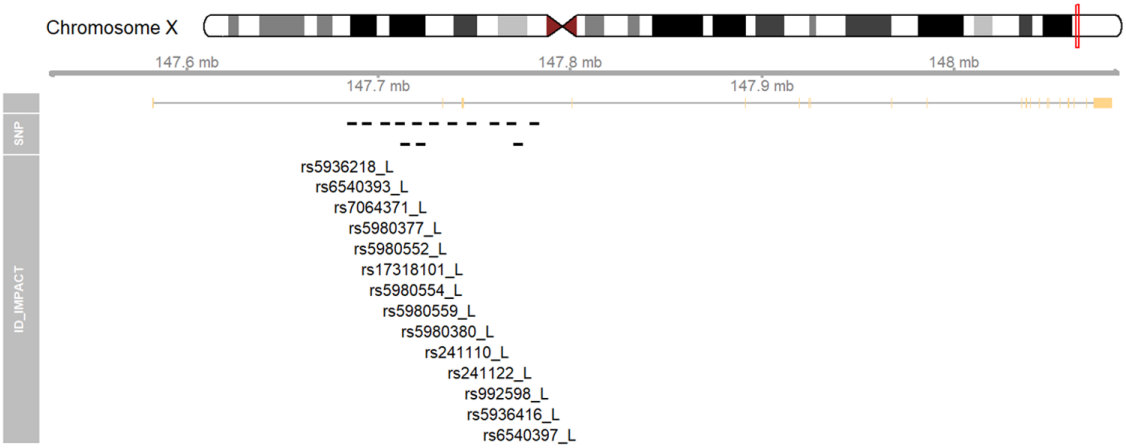


Figure 39. AFF2 gene map of bicluster 14.4. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

Bicluster 14.4 - Genic SNPs											
	Gene info.			SNP - Ensembl info.			SNP - Bicluster 14.4 info.				
	Gene	Rel	DEG	chrom	conseq_type	impact	%_NA	%_g1	%_g2	%_g3	Calc_MAF
rs17318101	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	36.84	0.00	63.16	0.63
rs241110	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	36.84	0.00	63.16	0.63
rs241122	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	15.79	0.00	84.21	0.84
rs5936218	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	15.79	0.00	84.21	0.84
rs5936416	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	31.58	0.00	68.42	0.68
rs5980377	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	15.79	0.00	84.21	0.84
rs5980380	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	36.84	0.00	63.16	0.63
rs5980552	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	36.84	0.00	63.16	0.63
rs5980554	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	15.79	0.00	84.21	0.84
rs5980559	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	36.84	0.00	63.16	0.63
rs6540393	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	15.79	0.00	84.21	0.84
rs6540397	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	42.11	0.00	57.89	0.58
rs7064371	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	15.79	0.00	84.21	0.84
rs992598	AFF2	IND.CRC	DOWN	X	intron_variant	LOW	0	36.84	0.00	63.16	0.63

Figure 40. Segment of the summary table of genic SNPs in bicluster 14.4 displaying the SNPs located in the AFF2 gene.

This gene encodes a putative **transcriptional activator** [103] and functions as an **RNA-binding protein (RBP)**, suggesting its potential involvement in the regulation of **alternative splicing** [104] [39]. In a manual search, the only evidence we have found linking this gene to CRC is from a study which identified a correlation between many significant alternative splicing events in CRC and specific RBPs, including AFF2. This outcome underscores a potential role for AFF2 in the observed **regulation of transcripts within CRC** [106]. Nevertheless, the limited external evidence regarding the involvement of this gene in the disease contrasts with the evidence suggested by our findings.

Interesting genes summary: we highlight PTCHD1-AS and AFF2 as genes whose SNPs appear crucial for CRC-bicluster identification. Despite limited external evidence of their involvement in CRC, the fact that these genes are affected in the majority of biclusters suggests that they may contribute to common clinical manifestations in a substantial portion of our pa-

[39] Alternative splicing is the process by which different combinations of gene segments (exons) lead to the production of various (alternative) mRNA strands, thereby diversifying a gene's protein production [105].

tient sample. This hypothesis is supported by the results of our RNA-Seq analysis, indicating PTCHD1 and AFF2 underexpression in CRC. Nevertheless, we also observe differences between biclusters; for instance, in bicluster 16.12, AFF2 is not affected, while a PTCHD1 SNP exhibits a MAF of 1, thus pointing towards new insights into CRC subtypes.

Appendix C.4 presents more examples of interesting genes, with data suggesting that their SNPs could influence their role in CRC.

3.7 Genotypic Cluster-Heatmap of Biclusters 2.2 and 8.4.

Figure 41 displays a **small portion** of a genotypic cluster-heatmap created with **genic SNPs** (in rows) and **patients** (in columns) from biclusters 2.2 and 8.4 (the pair that includes the highest proportion of patients). To enhance interpretability, both SNPs and patients have multiple annotations. SNP information is presented in the row names, and for patients, it is represented by color bars associated with values.

In a heatmap of such dimension, **identifying characteristic patterns of a single bicluster becomes challenging**, as each bicluster's characterization likely resides in the accumulation of subtle local differences. In other words, with this visualization, determining why PGMRA has included certain patients or SNPs in one bicluster, the other, or both is not straightforward.

Regarding the **patients' sex**, some differences are noticeable. Since women are the only ones who can be heterozygous (genotype *2*) for X chromosome SNPs that do not fall within pseudoautosomal regions, we can observe, for all SNPs corresponding to the gene ENSG00000155966 [40] (**first rows**), that the only **heterozygous genotypes** (represented with medium-intensity blue) belong to **female patients** (identified with yellow in the first color bar).

[40] This Ensembl identifier corresponds, precisely, to AFF2 gene.

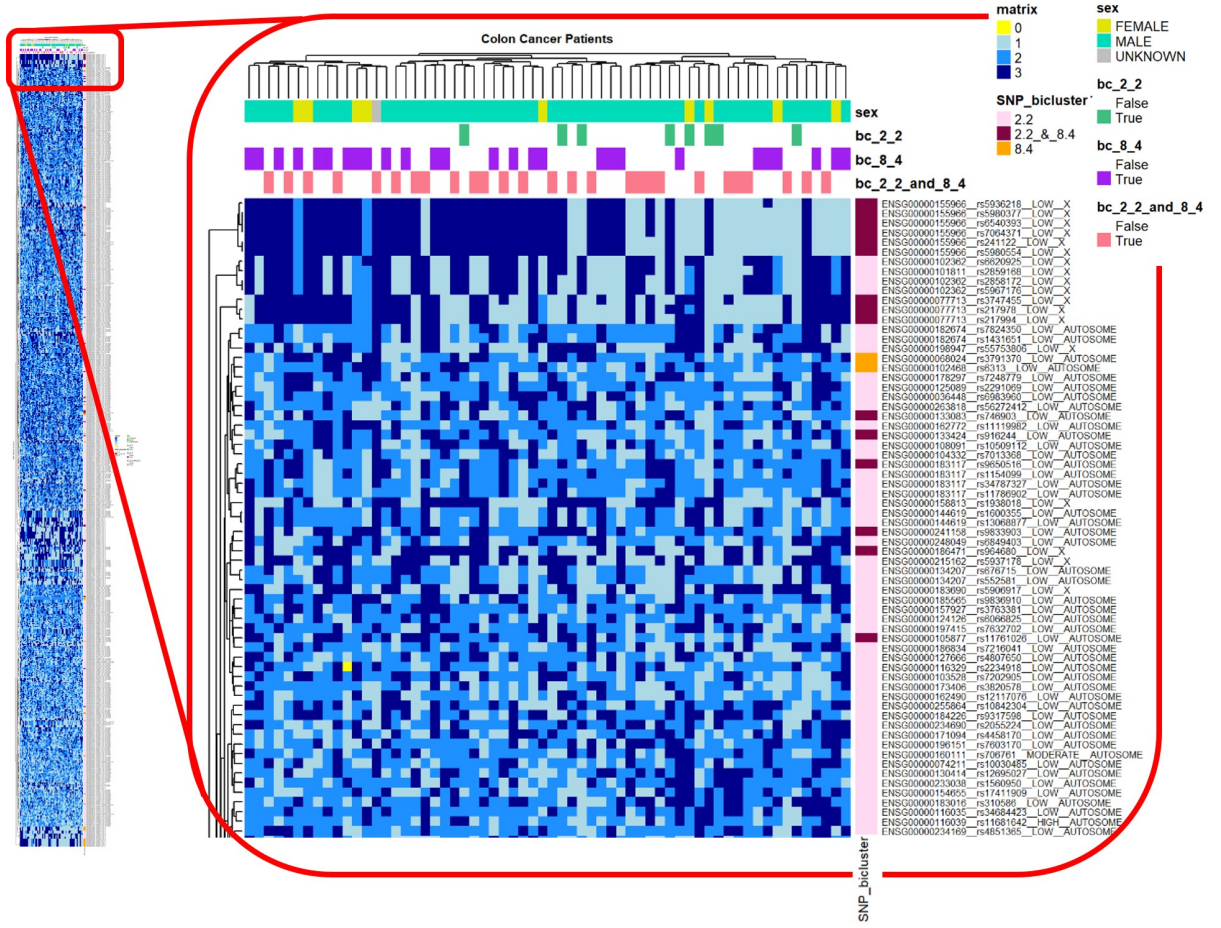


Figure 41. Portion of a long-format genotypic heatmap for biclusters 2.2 and 8.4. The row (SNP) names follow the structure: "Gene_SNP_Impact_Chromosome(X/Y/AUTOSOME)". The heatmap includes annotations denoting the membership of SNPs and patients in biclusters 2.2 and 8.4. Additionally, the sex of each patient is indicated. The arrangement of both rows and columns is organized using the "ward.D2" method to construct a dendrogram through hierarchical clustering.

Furthermore, it is noteworthy that we can easily identify rows corresponding to a specific gene. In this cluster-heatmap, the "ward.D2" method was employed to conduct hierarchical clustering on rows and columns, altering their order of appearance. As a consequence, we observe **clusters of highly similar SNPs**, resulting in **consecutive rows** in the heatmap. Considering the row labels, it becomes evident that the majority of these SNP clusters align with SNPs within the same gene. This observation implies that when a patient has any of the 3 genotypes for a SNP, they are likely to have the **same genotype** for the rest of the SNPs falling within the **same gene**. Visually, this translates into the appearance of vertical lines of a single color in the heatmap.

This intriguing pattern of consistent genotypes across multiple SNPs within a gene might be attributed to the concept of **linkage disequilibrium (LD)** (nonrandom association of alleles at two or more loci). LD introduces the likelihood of closely linked genetic variants being inherited together [107], constituting **haplotypes** [41], which could explain the absence of patients having different genotypes for the SNPs of the same gene. As an **illustrative example**, consider a hypothetical scenario where, for a specific genic SNP, a patient inherits the minor allele from their father and the major allele from their mother. This pattern is not unique to this particular SNP but extends to all SNPs belonging to the same gene, resulting in the patient having a consistent heterozygous genotype (*2*) across all SNPs of the gene. Some of the **genes that present haplotypes** in the cluster-heatmap (the complete one, not only the portion shown in the figure) can be found listed in Appendix C.5, Table 18.

[41] A haplotype (haploid genotype) is a group of alleles in an organism that are inherited together from a single parent [108].

CHAPTER 4: CONCLUSIONS AND FUTURE WORK

- This project involves utilizing the **PGMRA unsupervised algorithm** to identify **fuzzy CRC-biclusters (CRC patients x SNPs)** from genotypic data obtained from patients with CRC, breast cancer, skin cancer, and controls.
- Together, the 16 identified CRC-biclusters encompassed **73** CRC patients and **1997** SNPs, of which **1025** were intergenic and **972** were genic, affecting **695** different genes. The **global analysis (independent of bicluster-membership)** allowed:
  - The identification of **general clinical characteristics** in the entire sample: (e.g. low representation of patients under 50 years old and females) and **co-ocurrence trends** (e.g. basal CEA levels higher than 10 correlate with metastasis).
  - The identification of **general features of genic SNPs**: most of them are located on chromosome X, they generally affect protein-coding genes, and the most frequent transcript-level variant is intronic, associated with a low protein impact.
  - The **functional annotation of the 695 affected genes**, finding a direct relationship with CRC for 146 of them and an indirect relationship for 414. Over-representation analyses identified terms related to neurological disorders in the total gene set and additional terms linked to Ca2+ regulation in genes directly related to CRC. Both neurological disorders and hypercalcemia are known to be highly prevalent in CRC patients. These prevalences could have their roots in some of the genic SNPs on which PGMRA has relied to identify CRC-biclusters.
  - The identification of **differential expression in CRC for 72 of the 695 affected genes** through external RNA-Seq data, suggesting a role for these genes in the disease.
  - The creation of a **CRC protein-protein interaction minimum network**. The execution of the Louvain algorithm resulted in the partitioning of the network into 4 communities, with a modularity of 0.388. Different **over-represented terms** were identified for each community. In particular, the one with the highest proportion of seed nodes exhibited over-representation of the synapse term.
  - The **identification of 47 intergenic SNPs** that, according to RegulomeDB, have a minimum probability of 0.25 of exerting a **regulatory function in any of the specified organs**—colon, large intestine, intestine. All of these SNPs exhibit a very high

likelihood of possessing a general regulatory function, likely enhancing the probability of exerting specific functions in various tissues.

- The **analysis of patients and SNPs from each bicluster separately** allowed:
  - The **comparison of results** obtained across biclusters and with the global analysis.
  - The identification of **a particularly interesting bicluster, 16.12**, which includes a small number of patients and genic SNPs, characterized by generally high MAFs (minor allele frequencies). This bicluster exhibits the lowest average age (44.25). Observed gene functions related to inflammation and actin reorganization point to classify its members with the CRC **Consensus Molecular Subtype 4 (CMS4 - Mesenchymal)**. Additionally, these same CRC cases could be associated to **Inflammatory Bowel Disease**, explaining the low average age.
- We identified several genes whose SNPs appear crucial for CRC-bicluster identification, suggesting potential contributions to the disease, despite limited external evidence of their involvement. We specifically highlight **AFF2**, which encodes an RNA-binding protein, and **PTCHD1-AS**, which encodes a lncRNA. Both appear frequently in different biclusters, affected by a high number of SNPs and remarkably high MAFs in some cases. These combinations vary across biclusters. Moreover, our RNA-Seq analysis identified both AFF2 and the target gene of PTCHD1-AS as significantly underregulated genes in CRC.

Taking everything into account, **PGMRA results on small sample sizes showed the ability to uncover novel knowledge in the genetic characterization of CRC patients.** The results obtained from the complete sets of patients and SNPs pointed towards the genetic origins of medical conditions already known to be prevalent in CRC patients. On the other hand, the analysis of PGMRA biclusters separately revealed new insights into the complex interplay between genetic variants, clinical characteristics, and potential new perspectives on CRC subtypes that must be tested in the laboratory.

Future lines of research stemming from this work would involve:

- Studying intergenic SNPs falling within **enhancers or silencers** to gain insights into their regulatory roles.
- Conducting a **comparative analysis of MAFs** in the biclusters against a reference genome where the minor allele coincides, to assess the significance of deviations from the

reference.

- Studying how **SNPs located on the sex chromosomes** contribute to the clinical and molecular differences observed in CRC between sexes.

CHAPTER 5. BIBLIOGRAPHY

[1] J. D. WATSON and F. H. C. CRICK. "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid". In: *Nature* 171.4356 (Apr. 1953), pp. 737–738. DOI: 10.1038/171737a0. URL: https://doi.org/10.1038/171737a0.

[2] Stanislau Yatskevich, James Rhodes, and Kim Nasmyth. "Organization of Chromosomal DNA by SMC Complexes". In: *Annual Review of Genetics* 53.1 (Dec. 2019), pp. 445–482. DOI: 10.1146/annurev-genet-112618-043633. URL: https://doi.org/10.1146/annurev-genet-112618-043633.

[3] *DNA vs Genes vs Chromosomes: An Overview — my.clevelandclinic.org*. https://my.clevelandclinic.org/health/body/23064-dna-genes--chromosomes. [Accessed 20-11-2023].

[4] M.S. Jurica and G.A. Roybal. "RNA Splicing". In: *Encyclopedia of Biological Chemistry*. Elsevier, 2013, pp. 185–190. DOI: 10.1016/b978-0-12-378630-2.00674-5. URL: http://dx.doi.org/10.1016/B978-0-12-378630-2.00674-5.

[5] FRANCIS CRICK. "Central Dogma of Molecular Biology". In: *Nature* 227.5258 (Aug. 1970), pp. 561–563. DOI: 10.1038/227561a0. URL: https://doi.org/10.1038/227561a0.

[6] Francesca Nadalin. "Paired is better: local assembly algorithms for NGS paired reads and applications to RNA-Seq". PhD thesis. May 2014.

[7] Alexander F. Palazzo and Eliza S. Lee. "Non-coding RNA: what is functional and what is junk?" In: *Frontiers in Genetics* 6 (Jan. 2015). DOI: 10.3389/fgene.2015.00002. URL: https://doi.org/10.3389/fgene.2015.00002.

[8] Luca Del Giacco and Cristina Cattaneo. "Introduction to Genomics". In: *Methods in Molecular Biology*. Humana Press, Oct. 2011, pp. 79–88. DOI: 10.1007/978-1-60327-216-2_6. URL: https://doi.org/10.1007/978-1-60327-216-2_6.

[9] Virginia Espina, ed. *Molecular Profiling*. Springer New York, 2017. DOI: 10.1007/978-1-4939-6990-6. URL: https://doi.org/10.1007/978-1-4939-6990-6.

[10] Alan F Wright. *Genetic Variation: Polymorphisms and Mutations*. Sept. 2005. DOI: 10.1038/npg.els.0005005. URL: https://doi.org/10.1038/npg.els.0005005.

[11] lesliefischerrd. *What is a SNP? — nutrigeneticsspecialists.com*. https://www.nutrigeneticsspecialists.com/single-post/2017/03/27/what-is-a-snp. [Accessed 20-11-2023].

[12] Barkur S. Shastry. "SNPs: Impact on Gene Function and Phenotype". In: *Methods in Molecular Biology*. Humana Press, 2009, pp. 3–22. DOI: 10.1007/978-1-60327-411-1_1. URL: https://doi.org/10.1007/978-1-60327-411-1_1.

[13] Denise Zickler and Nancy Kleckner. "Recombination, Pairing, and Synapsis of Homologs during Meiosis". In: *Cold Spring Harbor Perspectives in Biology* 7.6 (May 2015), a016626. DOI: 10.1101/cshperspect.a016626. URL: https://doi.org/10.1101/cshperspect.a016626.

[14] *Allele frequency - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Allele_frequency. [Accessed 11-09-2023].

[15] *Frecuencia del alelo menos comun - Wikipedia - Wikipedia, la enciclopedia libre — es.wikipedia.org*. https://acortar.link/0TutTZ. [Accessed 20-11-2023].

[16] Eleanor Lawrence. *Hendersons dictionary of biology*. Pearson Education Limited, 2016.

[17] *Hemizygosity — biologyonline.com*. https://www.biologyonline.com/dictionary/hemizygosity. [Accessed 06-09-2023].

[18] Rüdiger Jörg Blaschke and Gudrun Rappold. "The pseudoautosomal regions, SHOX and disease". In: *Current Opinion in Genetics &amp Development* 16.3 (June 2006), pp. 233–239. DOI: 10.1016/j.gde.2006.04.004. URL: https://doi.org/10.1016/j.gde.2006.04.004.

[19] Feifei Zhao, Manshu Song, Youxin Wang, and Wei Wang. "Genetic model". In: *Journal of Cellular and Molecular Medicine* 20.4 (Jan. 2016), pp. 765–765. DOI: 10.1111/jcmm.12751. URL: https://doi.org/10.1111/jcmm.12751.

[20] John M Henshall, Rachel J Hawken, Sonja Dominik, and William Barendse. "Estimating the effect of SNP genotype on quantitative traits from pooled DNA samples". In: *Genetics Selection Evolution* 44.1 (Apr. 2012). DOI: 10.1186/1297-9686-44-12. URL: https://doi.org/10.1186/1297-9686-44-12.

[21] Geoffrey S. Ginsburg and Kathryn A. Phillips. "Precision Medicine: From Science To Value". In: *Health Affairs* 37.5 (May 2018), pp. 694–701. DOI: 10.1377/hlthaff.2017.1624. URL: https://doi.org/10.1377/hlthaff.2017.1624.

[22] Roddy Walsh, Sean J. Jurgens, Jeanette Erdmann, and Connie R. Bezzina. "Genome-wide association studies of cardiovascular disease". In: *Physiological Reviews* 103.3 (July 2023), pp. 2039–2055. DOI: 10.1152/physrev.00024.2022. URL: https://doi.org/10.1152/physrev.00024.2022.

[23] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, Michael E Goddard, and Peter M Visscher. "Common SNPs explain a large proportion of the heritability for human height". In: *Nature Genetics* 42.7 (June 2010), pp. 565–569. DOI: 10.1038/ng.608. URL: https://doi.org/10.1038/ng.608.

[24] J. Arnedo, C. del Val, G. A. de Erausquin, R. Romero-Zaliz, D. Svrakic, C. R. Cloninger, and I. Zwir. "PGMRA: a web server for (phenotype x genotype) many-to-many relation analysis in GWAS". In: *Nucleic Acids Research* 41.W1 (June 2013), W142–W149. DOI: 10.1093/nar/gkt496. URL: https://doi.org/10.1093/nar/gkt496.

[25] Chris Ding, Xiaofeng He, and Horst D. Simon. "On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering". In: *Proceedings of the 2005 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Jan. 2005. DOI: 10.1137/1.9781611972757.70. URL: https://doi.org/10.1137/1.9781611972757.70.

[26] *Non-negative matrix factorization - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Non-negative_matrix_factorization. [Accessed 07-09-2023].

[27] Daniel Lee and H. Sebastian Seung. "Algorithms for Non-negative Matrix Factorization". In: *Advances in Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000. URL: https://proceedings.neurips.cc/paper_files/paper/2000/file/f9d1152547c0bde01830b7e8bd60024c-Paper.pdf.

[28] Ana L.N. Fred and Anil K. Jain. "Combining multiple clusterings using evidence accumulation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27.6 (June 2005), pp. 835–850. DOI: 10.1109/tpami.2005.113. URL: https://doi.org/10.1109/tpami.2005.113.

[29] Ahmed Malki, Rasha Abu ElRuz, Ishita Gupta, Asma Allouch, Semir Vranic, and Ala-Eddin Al Moustafa. "Molecular Mechanisms of Colon Cancer Progression and Metastasis: Recent Insights and Advancements". In: *International Journal of Molecular Sciences* 22.1 (Dec. 2020), p. 130. DOI: 10.3390/ijms22010130. URL: https://doi.org/10.3390/ijms22010130.

[30] *Colorectal cancer - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Colorectal_cancer. [Accessed 20-11-2023].

[31] Evelien Dekker, Pieter J Tanis, Jasper L A Vleugels, Pashtoon M Kasi, and Michael B Wallace. "Colorectal cancer". In: *The Lancet* 394.10207 (Oct. 2019), pp. 1467–1480. DOI: 10.1016/s0140-6736(19)32319-0. URL: https://doi.org/10.1016/s0140-6736(19)32319-0.

[32] Fredrick R. Schumacher et al. "Genome-wide association study of colorectal cancer identifies six new susceptibility loci". In: *Nature Communications* 6.1 (July 2015). DOI: 10.1038/ncomms8138. URL: https://doi.org/10.1038/ncomms8138.

[33] Luuk Harbers, Federico Agostini, Marcin Nicos, Dimitri Poddighe, Magda Bienko, and Nicola Crosetto. "Somatic Copy Number Alterations in Human Cancers: An Analysis of Publicly Available Data From The Cancer Genome Atlas". In: *Frontiers in Oncology* 11 (July 2021). DOI: 10.3389/fonc.2021.700568. URL: https://doi.org/10.3389/fonc.2021.700568.

[34] Emma M. Schatoff, Benjamin I. Leach, and Lukas E. Dow. "WNT Signaling and Colorectal Cancer". In: *Current Colorectal Cancer Reports* 13.2 (Feb. 2017), pp. 101–110. DOI: 10.1007/s11888-017-0354-9. URL: https://doi.org/10.1007/s11888-017-0354-9.

[35] Hans Clevers and Roel Nusse. "Wnt/-Catenin Signaling and Disease". In: *Cell* 149.6 (June 2012), pp. 1192–1205. DOI: 10.1016/j.cell.2012.05.012. URL: https://doi.org/10.1016/j.cell.2012.05.012.

[36] Justin Guinney, Rodrigo Dienstmann, Xin Wang, Aurélien de Reyniès, Andreas Schlicker, Charlotte Soneson, Laetitia Marisa, Paul Roepman, Gift Nyamundanda, Paolo Angelino, Brian M Bot, Jeffrey S Morris, Iris M Simon, Sarah Gerster, Evelyn Fessler, Felipe De Sousa E Melo, Edoardo Missiaglia, Hena Ramay, David Barras, Krisztian Homicsko, Dipen Maru, Ganiraju C Manyam, Bradley Broom, Valerie Boige, Beatriz Perez-Villamil, Ted Laderas, Ramon Salazar, Joe W Gray, Douglas Hanahan, Josep Tabernero, Rene Bernards, Stephen H Friend, Pierre Laurent-Puig, Jan Paul Medema, Anguraj Sadanandam, Lodewyk Wessels, Mauro Delorenzi, Scott Kopetz, Louis Vermeulen, and Sabine Tejpar. "The consensus molecular subtypes of colorectal cancer". In: *Nature Medicine* 21.11 (Oct. 2015), pp. 1350–1356. DOI: 10.1038/nm.3967. URL: https://doi.org/10.1038/nm.3967.

[37] Alan White, Lucy Ironmonger, Robert J. C. Steele, Nick Ormiston-Smith, Carina Crawford, and Amanda Seims. "A review of sex-related differences in colorectal cancer incidence, screening uptake, routes to diagnosis, cancer stage and survival in the UK". In: *BMC Cancer* 18.1 (Sept. 2018). DOI: 10.1186/s12885-018-4786-7. URL: https://doi.org/10.1186/s12885-018-4786-7.

[38] Gwen Murphy, Susan S. Devesa, Amanda J. Cross, Peter D. Inskip, Katherine A. McGlynn, and Michael B. Cook. "Sex disparities in colorectal cancer incidence by anatomic subsite, race and age". In: *International Journal of Cancer* 128.7 (May 2010), pp. 1668–1675. DOI: 10.1002/ijc.25481. URL: https://doi.org/10.1002/ijc.25481.

[39] Patricia S. Steeg. "Targeting metastasis". In: *Nature Reviews Cancer* 16.4 (Mar. 2016), pp. 201–218. DOI: 10.1038/nrc.2016.25. URL: https://doi.org/10.1038/nrc.2016.25.

[40] Stefanie Gerstberger, Qingwen Jiang, and Karuna Ganesh. "Metastasis". In: *Cell* 186.8 (Apr. 2023), pp. 1564–1579. DOI: 10.1016/j.cell.2023.03.003. URL: https://doi.org/10.1016/j.cell.2023.03.003.

[41] Daniele V. F. Tauriello, Sergio Palomo-Ponce, Diana Stork, Antonio Berenguer-Llergo, Jordi Badia-Ramentol, Mar Iglesias, Marta Sevillano, Sales Ibiza, Adrià Cañellas, Xavier Hernando-Momblona, Daniel Byrom, Joan A. Matarin, Alexandre Calon, Elisa I. Rivas, Angel R. Nebreda, Antoni Riera, Camille Stephan-Otto Attolini, and Eduard Batlle. "TGF drives immune evasion in genetically reconstituted colon cancer metastasis". In: *Nature* 554.7693 (Feb. 2018), pp. 538–543. DOI: 10.1038/nature25492. URL: https://doi.org/10.1038/nature25492.

[42] Juan A. Marchal, Gabriel J. Lopez, Macarena Peran, Ana Comino, Juan R. Delgado, Javier A. García-García, Veronica Conde, Fernando M. Aranda, Carmen Rivas, Mariano Esteban, and Maria A. Garcia. "The impact of PKR activation: from neurodegeneration to cancer". In: *The FASEB Journal* 28.5 (Feb. 2014), pp. 1965–1974. DOI: 10.1096/fj.13-248294. URL: https://doi.org/10.1096/fj.13-248294.

[43] *Ensembl genome browser 110 — ensembl.org*. https://www.ensembl.org/index.html. [Accessed 21-11-2023].

[44] *Calculated consequences — ensembl.org*. https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html. [Accessed 12-11-2023].

[45] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. URL: http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

[46] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. "Mining association rules between sets of items in large databases". In: *ACM SIGMOD Record* 22.2 (June 1993), pp. 207–216. DOI: 10.1145/170036.170072. URL: https://doi.org/10.1145/170036.170072.

[47] Rakesh Agrawal and Ramakrishnan Srikant. "Fast Algorithms for Mining Association Rules in Large Databases". In: *Proceedings of the 20th International Conference on Very Large Data Bases*. VLDB '94. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1994, pp. 487–499. ISBN: 1558601538.

[48] Michael Hahsler, Bettina Grün, and Kurt Hornik. "barules/b- A Computational Environment for Mining Association Rules and Frequent Item Sets". In: *Journal of Statistical Software* 14.15 (2005). DOI: 10.18637/jss.v014.i15. URL: https://doi.org/10.18637/jss.v014.i15.

[49] Shengcheng Dong, Nanxiang Zhao, Emma Spragins, Meenakshi S. Kagda, Mingjie Li, Pedro Assis, Otto Jolanki, Yunhai Luo, J. Michael Cherry, Alan P. Boyle, and Benjamin C. Hitz. "Annotating and prioritizing human non-coding variants with RegulomeDB v.2". In: *Nature Genetics* 55.5 (Apr. 2023), pp. 724–726. DOI: 10.1038/s41588-023-01365-3. URL: https://doi.org/10.1038/s41588-023-01365-3.

[50] J. Pinero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong. "DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes". In: *Database* 2015.0 (Apr. 2015), bav028–bav028. DOI: 10.1093/database/bav028. URL: https://doi.org/10.1093/database/bav028.

[51] Guangchuang Yu. *Introduction — Biomedical Knowledge Mining using GOSemSim and clusterProfiler — yulab-smu.top*. https://yulab-smu.top/biomedical-knowledge-mining-book/index.html. [Accessed 03-11-2023].

[52] *Gene Ontology Resource — geneontology.org*. https://www.geneontology.org/. [Accessed 03-11-2023].

[53] *KEGG: Kyoto Encyclopedia of Genes and Genomes — genome.jp*. https://www.genome.jp/kegg/. [Accessed 03-11-2023].

[54] Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. "GO::TermFinder—open source software for accessing Gene Ontology informa-

tion and finding significantly enriched Gene Ontology terms associated with a list of genes". In: *Bioinformatics* 20.18 (Aug. 2004), pp. 3710–3715. DOI: 10.1093/bioinformatics/bth456. URL: https://doi.org/10.1093/bioinformatics/bth456.

[55] Elizabeth I. Boyle, Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, and Gavin Sherlock. "GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". In: *Bioinformatics* 20.18 (Aug. 2004), pp. 3710–3715. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/bth456. URL: http://dx.doi.org/10.1093/bioinformatics/bth456.

[56] Ashraful Haque, Jessica Engel, Sarah A. Teichmann, and Tapio Lönnberg. "A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications". In: *Genome Medicine* 9.1 (Aug. 2017). ISSN: 1756-994X. DOI: 10.1186/s13073-017-0467-4. URL: http://dx.doi.org/10.1186/s13073-017-0467-4.

[57] Daniel Castillo-Secilla, Juan Manuel Gálvez, Francisco Carrillo-Perez, Marta Verona-Almeida, Daniel Redondo-Sánchez, Francisco Manuel Ortuno, Luis Javier Herrera, and Ignacio Rojas. "KnowSeq R-Bioc package: The automatic smart gene expression tool for retrieving relevant biological knowledge". In: *Computers in Biology and Medicine* 133 (June 2021), p. 104387. DOI: 10.1016/j.compbiomed.2021.104387. URL: https://doi.org/10.1016/j.compbiomed.2021.104387.

[58] Tianwen Wang, Ningning Yang, Chen Liang, Hongjv Xu, Yafei An, Sha Xiao, Mengyuan Zheng, Lu Liu, Gaozhan Wang, and Lei Nie. "Detecting Protein-Protein Interaction Based on Protein Fragment Complementation Assay". In: *Current Protein &amp Peptide Science* 21.6 (Aug. 2020), pp. 598–610. DOI: 10.2174/1389203721666200213102829. URL: https://doi.org/10.2174/1389203721666200213102829.

[59] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks". In: *Genome Research* 13.11 (Nov. 2003), pp. 2498–2504. ISSN: 1088-9051. DOI: 10.1101/gr.1239303. URL: http://dx.doi.org/10.1101/gr.1239303.

[60] Fuquan Zhang. "A flexible tool to plot a genomic map for single nucleotide polymorphisms". In: *Source Code for Biology and Medicine* 11.1 (Apr. 2016). ISSN: 1751-0473. DOI: 10.1186/s13029-016-0052-z. URL: http://dx.doi.org/10.1186/s13029-016-0052-z.

[61] URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=PCDH11X.

[62] Joo Han Lee and Seong-Wook Lee. "The Roles of Carcinoembryonic Antigen in Liver Metastasis and Therapeutic Approaches". In: *Gastroenterology Research and Practice* 2017 (2017), pp. 1–11.

ISSN: 1687-630X. DOI: 10.1155/2017/7521987. URL: http://dx.doi.org/10.1155/2017/7521987.

[63] Shane Lloyd, David Baraghoshi, Randa Tao, Ignacio Garrido-Laguna, Glynn W. Gilcrease, Jonathan Whisenant, John R. Weis, Courtney Scaife, Thomas B. Pickron, Lyen C. Huang, Marcus M. Monroe, Sarah Abdelaziz, Alison M. Fraser, Ken R. Smith, Vikrant Deshmukh, Michael Newman, Kerry G. Rowe, John Snyder, Niloy J. Samadder, and Mia Hashibe. "Mental Health Disorders are More Common in Colorectal Cancer Survivors and Associated With Decreased Overall Survival". In: *American Journal of Clinical Oncology* 42.4 (Apr. 2019), pp. 355–362. DOI: 10.1097/coc.0000000000000529. URL: https://doi.org/10.1097/coc.0000000000000529.

[64] Jessica A. Boyette-Davis, Cathy Eng, Xin S. Wang, Charles S. Cleeland, Gwen Wendelschafer-Crabb, William R. Kennedy, Donald A. Simone, Haijun Zhang, and Patrick M. Dougherty. "Subclinical Peripheral Neuropathy Is a Common Finding in Colorectal Cancer Patients Prior to Chemotherapy". In: *Clinical Cancer Research* 18.11 (May 2012), pp. 3180–3187. DOI: 10.1158/1078-0432.ccr-12-0205. URL: https://doi.org/10.1158/1078-0432.ccr-12-0205.

[65] Simone L. Schonkeren, Meike S. Thijssen, Nathalie Vaes, Werend Boesmans, and Veerle Melotte. "The Emerging Role of Nerves and Glia in Colorectal Cancer". In: *Cancers* 13.1 (Jan. 2021), p. 152. ISSN: 2072-6694. DOI: 10.3390/cancers13010152. URL: http://dx.doi.org/10.3390/cancers13010152.

[66] P K Vogt. "Cancer genes". en. In: *West. J. Med.* 158.3 (Mar. 1993), pp. 273–278.

[67] Wei Wang, Suyun Yu, Shuai Huang, Rui Deng, Yushi Ding, Yuanyuan Wu, Xiaoman Li, Aiyun Wang, Shijun Wang, Wenxing Chen, and Yin Lu. "A Complex Role for Calcium Signaling in Colorectal Cancer Development and Progression". In: *Molecular Cancer Research* 17.11 (Nov. 2019), pp. 2145–2153. DOI: 10.1158/1541-7786.mcr-19-0429. URL: https://doi.org/10.1158/1541-7786.mcr-19-0429.

[68] Dianne S. Schwarz and Michael D. Blower. "The endoplasmic reticulum: structure, function and response to cellular signaling". In: *Cellular and Molecular Life Sciences* 73.1 (Oct. 2015), pp. 79–94. ISSN: 1420-9071. DOI: 10.1007/s00018-015-2052-6. URL: http://dx.doi.org/10.1007/s00018-015-2052-6.

[69] Jessica Plati, Octavian Bucur, and Roya Khosravi-Far. "Dysregulation of apoptotic signaling in cancer: Molecular mechanisms and therapeutic opportunities". In: *Journal of Cellular Biochemistry* 104.4 (May 2008), pp. 1124–1149. ISSN: 1097-4644. DOI: 10.1002/jcb.21707. URL: http://dx.doi.org/10.1002/jcb.21707.

[70] *Phosphatidylinositol signaling system - CUSABIO — cusabio.com*. https://www.cusabio.com/pathway/Phosphatidylinositol-signaling-system.html. [Accessed 13-11-2023].

[71] Eric R Gamazon, Heather E Wheeler, Kaanan P Shah, Sahar V Mozaffari, Keston Aquino-Michaels, Robert J Carroll, Anne E Eyler, Joshua C Denny, Dan L Nicolae, Nancy J Cox, and Hae Kyung Im. "A gene-based association method for mapping traits using reference transcriptome data". In: *Nature Genetics* 47.9 (Aug. 2015), pp. 1091–1098. ISSN: 1546-1718. DOI: 10.1038/ng.3367. URL: http://dx.doi.org/10.1038/ng.3367.

[72] Eddie Cano-Gamez and Gosia Trynka. "From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases". In: *Frontiers in Genetics* 11 (May 2020). ISSN: 1664-8021. DOI: 10.3389/fgene.2020.00424. URL: http://dx.doi.org/10.3389/fgene.2020.00424.

[73] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. "Finding community structure in very large networks". In: *Phys. Rev. E* 70 (6 Dec. 2004), p. 2. DOI: 10.1103/PhysRevE.70.066111. URL: https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

[74] Maria Mittelbrunn, Miguel Vicente-Manzanares, and Francisco Sánchez-Madrid. "Organizing Polarized Delivery of Exosomes at Synapses". In: *Traffic* 16.4 (Mar. 2015), pp. 327–337. ISSN: 1600-0854. DOI: 10.1111/tra.12258. URL: http://dx.doi.org/10.1111/tra.12258.

[75] Rong Fu, Xiaowan Jiang, Gang Li, Yi Zhu, and Huimin Zhang. "Junctional complexes in epithelial cells: sentinels for extracellular insults and intracellular homeostasis". In: *The FEBS Journal* 289.23 (Sept. 2021), pp. 7314–7333. ISSN: 1742-4658. DOI: 10.1111/febs.16174. URL: http://dx.doi.org/10.1111/febs.16174.

[76] C. Sidhanth, P. Manasa, S. Krishnapriya, S. Sneha, S. Bindhya, R.P. Nagare, M. Garg, and T.S. Ganesan. "A systematic understanding of signaling by ErbB2 in cancer using phosphoproteomics". In: *Biochemistry and Cell Biology* 96.3 (June 2018), pp. 295–305. DOI: 10.1139/bcb-2017-0020. URL: https://doi.org/10.1139/bcb-2017-0020.

[77] Apr. 2022. URL: https://www.cdc.gov/ibd/what-is-IBD.htm#:~:text=Inflammatory%20bowel%20disease%20(IBD)%20is,Close.

[78] Ziad Kanaan, Motaz Qadan, Maurice Robert Eichenberger, and Susan Galandiuk. "The Actin-Cytoskeleton Pathway and Its Potential Role in Inflammatory Bowel Disease-Associated Human Colorectal Cancer". In: *Genetic Testing and Molecular Biomarkers* 14.3 (June 2010), pp. 347–353. ISSN: 1945-0257. DOI: 10.1089/gtmb.2009.0197. URL: http://dx.doi.org/10.1089/gtmb.2009.0197.

[79] *Leiomyosarcoma — cancer.gov*. https://www.cancer.gov/pediatric-adult-rare-tumor/rare-tumors/rare-soft-tissue-tumors/leiomyosarcoma. [Accessed 15-11-2023].

[80] Fady Hannah-Shmouni, Giampaolo Trivellin, and Constantine A. Stratakis. "Genetics of gigantism and acromegaly". In: *Growth Hormone amp; IGF Research* 30–31 (Oct. 2016), pp. 37–41. ISSN:

1096-6374. DOI: 10.1016/j.ghir.2016.08.002. URL: http://dx.doi.org/10.1016/j.ghir.2016.08.002.

[81] *Gigantismo: MedlinePlus enciclopedia médica — medlineplus.gov*. https://medlineplus.gov/spanish/ency/article/001174.htm. [Accessed 15-11-2023].

[82] Stephan Meding, Benjamin Balluff, Mareike Elsner, Cédrik Schöne, Sandra Rauser, Ulrich Nitsche, Matthias Maak, Alexander Schäfer, Stefanie M Hauck, Marius Ueffing, Rupert Langer, Heinz Höfler, Helmut Friess, Robert Rosenberg, and Axel Walch. "Tissue-based proteomics reveals FXYD3, S100A11 and GSTM3 as novel markers for regional lymph node metastasis in colon cancer". In: *The Journal of Pathology* 228.4 (May 2012), pp. 459–470. ISSN: 1096-9896. DOI: 10.1002/path.4021. URL: http://dx.doi.org/10.1002/path.4021.

[83] *Cytoskeleton - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Cytoskeleton. [Accessed 29-11-2023].

[84] *Actin - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Actin. [Accessed 15-11-2023].

[85] George C. Prendergast and Jackson B. Gibbs. "Pathways of Ras Function: Connections to the Actin Cytoskeleton". In: *Advances in Cancer Research*. Elsevier, 1993, pp. 19–64. DOI: 10.1016/s0065-230x(08)60314-0. URL: https://doi.org/10.1016/s0065-230x(08)60314-0.

[86] Nahum Zepeta-Flores, Mahara Valverde, Alejandro Lopez-Saavedra, and Emilio Rojas. "Glutathione depletion triggers actin cytoskeleton changes via actin-binding proteins". In: *Genetics and Molecular Biology* 41.2 (June 2018), pp. 475–487. DOI: 10.1590/1678-4685-gmb-2017-0158. URL: https://doi.org/10.1590/1678-4685-gmb-2017-0158.

[87] Lara E. Davis, Kevin D. Nusser, Joanna Przybyl, Janét Pittsenbarger, Nicolle E. Hofmann, Sushama Varma, Sujay Vennam, Maria Debiec-Rychter, Matt van de Rijn, and Monika A. Davare. "Discovery and Characterization of Recurrent, Targetable ALK Fusions in Leiomyosarcoma". In: *Molecular Cancer Research* 17.3 (Mar. 2019), pp. 676–685. ISSN: 1557-3125. DOI: 10.1158/1541-7786.mcr-18-1075. URL: http://dx.doi.org/10.1158/1541-7786.MCR-18-1075.

[88] Ah Ra Jung, Chan-Hun Jung, Joo Kyung Noh, Young Chan Lee, and Young-Gyu Eun. "Epithelial-mesenchymal transition gene signature is associated with prognosis and tumor microenvironment in head and neck squamous cell carcinoma". In: *Scientific Reports* 10.1 (Feb. 2020). ISSN: 2045-2322. DOI: 10.1038/s41598-020-60707-x. URL: http://dx.doi.org/10.1038/s41598-020-60707-x.

[89] XUEBING YAN, LEILEI YAN, SIHONG LIU, ZEZHI SHAN, YUAN TIAN, and ZHIMING JIN. "N-cadherin, a novel prognostic biomarker, drives malignant progression of colorectal cancer". In: *Molecular Medicine Reports* 12.2 (Apr. 2015), pp. 2999–3006. ISSN: 1791-3004. DOI: 10.3892/mmr.2015.3687. URL: http://dx.doi.org/10.3892/mmr.2015.3687.

[90] Jay Shankar, Anat Messenberg, Jackie Chan, T. Michael Underhill, Leonard J. Foster, and Ivan R. Nabi. "Pseudopodial Actin Dynamics Control Epithelial-Mesenchymal Transition in Metastatic Cancer Cells". In: *Cancer Research* 70.9 (Apr. 2010), pp. 3780–3790. ISSN: 1538-7445. DOI: 10.1158/0008-5472.can-09-4439. URL: http://dx.doi.org/10.1158/0008-5472.CAN-09-4439.

[91] Dhanunjay Mukhi, Lakshmi P. Kolligundla, Saikrishna Maruvada, Rajkishor Nishad, and Anil K. Pasupulati. "Growth hormone induces transforming growth factor-1 in podocytes: Implications in podocytopathy and proteinuria". In: *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1870.2 (Feb. 2023), p. 119391. ISSN: 0167-4889. DOI: 10.1016/j.bbamcr.2022.119391. URL: http://dx.doi.org/10.1016/j.bbamcr.2022.119391.

[92] Omar Vergara-Fernandez, Mario Trejo-Avila, and Noel Salgado-Nesme. "Sarcopenia in patients with colorectal cancer: A comprehensive review". In: *World Journal of Clinical Cases* 8.7 (Apr. 2020), pp. 1188–1202. ISSN: 2307-8960. DOI: 10.12998/wjcc.v8.i7.1188. URL: http://dx.doi.org/10.12998/wjcc.v8.i7.1188.

[93] Sara Lovisa, Giannicola Genovese, and Silvio Danese. "Role of Epithelial-to-Mesenchymal Transition in Inflammatory Bowel Disease". In: *Journal of Crohn's and Colitis* 13.5 (Dec. 2018), pp. 659–668. ISSN: 1876-4479. DOI: 10.1093/ecco-jcc/jjy201. URL: http://dx.doi.org/10.1093/ecco-jcc/jjy201.

[94] Amritpal Dhaliwal, Jonathan I. Quinlan, Kellie Overthrow, Carolyn Greig, Janet M. Lord, Matthew J. Armstrong, and Sheldon C. Cooper. "Sarcopenia in Inflammatory Bowel Disease: A Narrative Overview". In: *Nutrients* 13.2 (Feb. 2021), p. 656. ISSN: 2072-6643. DOI: 10.3390/nu13020656. URL: http://dx.doi.org/10.3390/nu13020656.

[95] *GSS glutathione synthetase [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov*. https://www.ncbi.nlm.nih.gov/gene/2937#summary. [Accessed 17-11-2023].

[96] Taunia D. Lee, Heping Yang, Janet Whang, and Shelly C. Lu. "Cloning and characterization of the human glutathione synthetase 5-flanking region". In: *Biochemical Journal* 390.2 (Aug. 2005), pp. 521–528. ISSN: 1470-8728. DOI: 10.1042/bj20050439. URL: http://dx.doi.org/10.1042/BJ20050439.

[97] *DDOST dolichyl-diphosphooligosaccharide–protein glycosyltransferase non-catalytic subunit [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov*. https://www.ncbi.nlm.nih.gov/gene/1650#summary. [Accessed 17-11-2023].

[98] *STT3B STT3 oligosaccharyltransferase complex catalytic subunit B [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov*. https://www.ncbi.nlm.nih.gov/gene/201595#summary. [Accessed 17-11-2023].

[99] Vicent Pelechano and Lars M. Steinmetz. "Gene regulation by antisense transcription". In: *Nature Reviews Genetics* 14.12 (Nov. 2013), pp. 880–893. ISSN: 1471-0064. DOI: 10.1038/nrg3594. URL: http://dx.doi.org/10.1038/nrg3594.

[100] Wanjing Chen, Qian Tu, Liang Yu, Yanyan Xu, Gang Yu, Benli Jia, Yunsheng Cheng, and Yong Wang. "LncRNA ADAMTS9-AS1, as prognostic marker, promotes cell proliferation and EMT in colorectal cancer". In: *Human Cell* 33.4 (Sept. 2020), pp. 1133–1141. ISSN: 1749-0774. DOI: 10.1007/s13577-020-00388-w. URL: http://dx.doi.org/10.1007/s13577-020-00388-w.

[101] [Accessed 17-11-2023]. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=ADAMTS9.

[102] [Accessed 17-11-2023]. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=PTCHD1.

[103] *AFF2 ALF transcription elongation factor 2 [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov*. https://www.ncbi.nlm.nih.gov/gene/2334#summary. [Accessed 18-11-2023].

[104] *UniProt — uniprot.org*. https://www.uniprot.org/uniprotkb/P51816/entry#function. [Accessed 18-11-2023].

[105] *Alternative splicing - Wikipedia — en.wikipedia.org*. https://en.wikipedia.org/wiki/Alternative_splicing. [Accessed 08-12-2023].

[106] Kavitha Mukund, Natalia Syulyukina, Sonia Ramamoorthy, and Shankar Subramaniam. "Right and left-sided colon cancers - specificity of molecular mechanisms in tumorigenesis and progression". In: *BMC Cancer* 20.1 (Apr. 2020). ISSN: 1471-2407. DOI: 10.1186/s12885-020-06784-7. URL: http://dx.doi.org/10.1186/s12885-020-06784-7.

[107] Montgomery Slatkin. "Linkage disequilibrium — understanding the evolutionary past and mapping the medical future". In: *Nature Reviews Genetics* 9.6 (June 2008), pp. 477–485. ISSN: 1471-0064. DOI: 10.1038/nrg2361. URL: http://dx.doi.org/10.1038/nrg2361.

[108] Editorial Board. *Concise Dictionary of Science*. New Delhi, India: V & S, May 2012.

[109] Gongcheng Li, Tiejun Pan, Dan Guo, and Long-Cheng Li. "Regulatory Variants and Disease: The E-Cadherin -160C/A SNP as an Example". In: *Molecular Biology International* 2014 (Sept. 2014), pp. 1–9. DOI: 10.1155/2014/967565. URL: https://doi.org/10.1155/2014/967565.

[110] Chava Kimchi-Sarfaty, Jung Mi Oh, In-Wha Kim, Zuben E. Sauna, Anna Maria Calcagno, Suresh V. Ambudkar, and Michael M. Gottesman. "A "Silent" Polymorphism in the iMDR/i 1 Gene Changes Substrate Specificity". In: *Science* 315.5811 (Jan. 2007), pp. 525–528. DOI: 10.1126/science.1135308. URL: https://doi.org/10.1126/science.1135308.

[111] Mohammad Al-Haggar, Agnieszka Madej-Pilarczyk, Lukasz Kozlowski, Janusz M Bujnicki, Sohier Yahia, Dina Abdel-Hadi, Amany Shams, Nermin Ahmad, Sahar Hamed, and Monika Puzianowska-Kuznicka. "A novel homozygous p.Arg527Leu LMNA mutation in two unrelated Egyptian families causes overlapping mandibuloacral dysplasia and progeria syndrome". In: *European Journal of*

Human Genetics 20.11 (May 2012), pp. 1134–1140. DOI: 10.1038/ejhg.2012.77. URL: https://doi.org/10.1038/ejhg.2012.77.

[112] S.K. Cordovado, M. Hendrix, C.N. Greene, S. Mochal, M.C. Earley, P.M. Farrell, M. Kharrazi, W.H. Hannon, and P.W. Mueller. "CFTR mutation analysis and haplotype associations in CF patients". In: *Molecular Genetics and Metabolism* 105.2 (Feb. 2012), pp. 249–254. DOI: 10.1016/j.ymgme.2011.10.013. URL: https://doi.org/10.1016/j.ymgme.2011.10.013.

[113] *Illumina Microarray Technology — illumina.com*. https://www.illumina.com/science/technology/microarray.html. [Accessed 21-11-2023].

[114] Mark Newman. *Networks*. Oxford University Press, Mar. 2010. DOI: 10.1093/acprof:oso/9780199206650.001.0001. URL: https://doi.org/10.1093/acprof:oso/9780199206650.001.0001.

[115] Stanley Wasserman and Katherine Faust. *Análisis de redes sociales. Métodos y aplicaciones*. en. CIS- Centro de Investigaciones Sociológicas, Dec. 2013.

[116] Aaron Clauset, M. E. J. Newman, and Cristopher Moore. "Finding community structure in very large networks". In: *Physical Review E* 70.6 (Dec. 2004). DOI: 10.1103/physreve.70.066111. URL: https://doi.org/10.1103/physreve.70.066111.

[117] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. "Fast unfolding of communities in large networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 2008), P10008. DOI: 10.1088/1742-5468/2008/10/p10008. URL: https://doi.org/10.1088/1742-5468/2008/10/p10008.

[118] Anthony J F Griffiths. *Genética*. es. 2002.

[119] Douglas Hanahan. "Hallmarks of Cancer: New Dimensions". In: *Cancer Discovery* 12.1 (Jan. 2022), pp. 31–46. ISSN: 2159-8290. DOI: 10.1158/2159-8290.cd-21-1059. URL: http://dx.doi.org/10.1158/2159-8290.CD-21-1059.

[120] Moslem Bahadori. "New insights into connection of nucleolar functions and cancer". en. In: *Tanaffos* 18.3 (Mar. 2019), pp. 173–179.

[121] *Nerve Growth Factor Signaling Pathway - Creative Diagnostics — creative-diagnostics.com*. https://www.creative-diagnostics.com/nerve-growth-factor-signaling-pathway.htm. [Accessed 14-11-2023].

[122] *TNFRSF17 TNF receptor superfamily member 17 [Homo sapiens (human)] - Gene - NCBI — ncbi.nlm.nih.gov*. https://www.ncbi.nlm.nih.gov/gene/608#summary. [Accessed 17-11-2023].

[123] Soo-Cheon Chae, Ji-In Yu, Tai-Boong Uhm, Sam-Yun Lee, Dong-Baek Kang, Jeong-Kyun Lee, Won-Cheol Park, and Ki-Jung Yun. "The haplotypes of TNFRSF17 polymorphisms are associated with colon cancer in a Korean population". In: *International Journal of Colorectal Disease* 27.6

(Nov. 2011), pp. 701–707. ISSN: 1432-1262. DOI: 10.1007/s00384-011-1364-8. URL: http://dx.doi.org/10.1007/s00384-011-1364-8.

[124] Ine Jorgensen and Edward A. Miao. "Pyroptotic cell death defends against intracellular pathogens". In: *Immunological Reviews* 265.1 (Apr. 2015), pp. 130–142. ISSN: 1600-065X. DOI: 10.1111/imr.12287. URL: http://dx.doi.org/10.1111/imr.12287.

[125] Mingchao Mu, Qiaoling Yu, Qin Zhang, Jing Guo, Xingjie Wang, Xuejun Sun, and Junhui Yu. "A pan-cancer analysis of molecular characteristics and oncogenic role of gasdermins". In: *Cancer Cell International* 22.1 (Feb. 2022). ISSN: 1475-2867. DOI: 10.1186/s12935-022-02483-4. URL: http://dx.doi.org/10.1186/s12935-022-02483-4.

[126] [Accessed 17-11-2023]. URL: https://www.genecards.org/cgi-bin/carddisp.pl?gene=GSDMB.

[127] Niki Christou, Aurélie Perraud, Sabrina Blondy, Marie-Odile Jauberteau, Serge Battu, and Muriel Mathonnet. "E-cadherin: A potential biomarker of colorectal cancer prognosis". In: *Oncology Letters* 13.6 (Apr. 2017), pp. 4571–4576. ISSN: 1792-1082. DOI: 10.3892/ol.2017.6063. URL: http://dx.doi.org/10.3892/ol.2017.6063.

A Introduction - Supplementary Material.

A.1 Examples of how any Type of SNP Can Give Rise to a Observable Phenotype.

- Certain SNPs in **non-coding regions** correlate with an increased likelihood of developing cancer [109].
- Some SNPs in **coding regions**, consisting of **synonymous substitutions**, despite not altering the amino acid sequence of the protein, could still affect its function. This occurs, for instance, in the case of the Multidrug Resistance Protein 1 (MDR1) receptor, where a silent mutation SNP slows down the translation of the nascent peptide, causing it to fold into an alternative conformation less functional than the native three-dimensional structure [110].
- SNPs involving an **amino acid substitution** (protein change) are the most frequently associated with the onset of diseases. An example of this is the SNP 1580G>T in the LMNA gene, causing the change from arginine to leucine in the protein, a phenotype related to diseases such as progeria or mandibuloacral dysplasia [111].
- **Nonsense SNPs** lead to the appearance of a premature stop codon [42] that truncates the resulting protein, rendering it incomplete and usually non-functional. This is reflected in the G542X mutation in the CTFR gene, responsible for cystic fibrosis [112].

B Methodology - Supplementary Material.

B.1 Illumina Microarray Technology.

This method involves silica microbeads placed in microwells, coated with copies of a probe targeting a specific locus in the genome. As DNA fragments pass over the BeadChip, probes bind to sample DNA, and allele specificity is determined by a labeled nucleotide. The emitted signal, when excited by a laser, provides information about the allelic ratio at that locus [113].

[42] A start codon is a three-nucleotide sequence in mRNA that signal the beginning of protein synthesis and a stop codon is a three-nucleotide sequence that marks the end of protein synthesis.

B.2 Data Engineering Steps on the Clinical Dataset.

- **Column removal** for those with over 90% missing values or constant values across all rows.
- **Translation** (from Spanish to English) of the entire table (every attribute and all its categories).
- **Grouping of categories** when they meant the same thing (for example, in the "Histolog" attribute, categories like "ADENOCARCINOMA", "Adenocarcinoma ", "Adenocarcinoma" or "adenocarcinoma", were grouped into "adenocarcinoma").
- **Simplification** of some categorizations (for instance, in the "Stage" attribute, the categories "IIIA", "IIIB", and "IIIC", were merged into "III").
- **Discretization** of numerical variables (specifically, for the "Age" variable and the "basal_CEA" variable).
- Conversion of boolean columns to **categorical format** as well (changing values *0* and *1* to "No" and "Yes").
- Conversion of all missing values (empty strings and *NA* values [43]) to **"unavailable"**.
- Conversion of each column to the **"factor" type** to establish an order (levels) for the categories of each attribute. This was necessary to achieve the desired appearance order on the x-axis of bar charts.

B.3 Explanations of Some Terms Related to Graph Theory.

- **Node centrality** in a network refers to a metric that assesses a node's significance within the network based on its position and connections. This metric is crucial for understanding the network's structure and dynamics, as it pinpoints which nodes are the most influential or central in terms of communication, control, or information flow. There are various types of node centrality metrics [114].
- **Eigenvector centrality** is arguably the most commonly used centrality measure and is based on the concept that a node's centrality depends on the centrality of its neighbors. Therefore, having a single highly central neighbor is more significant than having many less central neighbors [115]. The calculation can be intricate due to its recursive nature.
- The **betweenness centrality** of a node is based on the number of shortest paths (geodesics)

[43] In R, a value is of the type *NA* when it is Not Available.

in the network that pass through that node. This metric takes into account the overall structure of the network and is focused on capturing brokerage. A node is more central when more pairs of nodes need to pass through it to indirectly connect with each other, making it a crucial element in linking different regions of the network [115].

B.4 Extraction of Association Rules: Brute-force Extraction, Two-step Approach and Apriori algorithm.

Brute-force rule extraction is a computationally prohibitive approach, involving the generation of all possible association rules, the calculation of their support and confidence, and filtering based on threshold metrics.

Similarly, the **two-step approach**, which consists of first generating **frequent itemsets** (those with support equal to or greater than the support threshold) and then generating high-confidence rules from frequent itemsets, is **computationally expensive**. In this approach, each rule is a binary partition of a frequent itemset, requiring the generation of all possible itemsets (2^d, where d is the number of items).

The **Apriori method** provides a feasible solution to the problem. This algorithm is based on the principle that if an itemset is frequent, then all its subsets are also frequent, maintaining this principle through the **anti-monotonic property of the support measure** [46]:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y) \tag{4}$$

The algorithm begins by generating all frequent itemsets of length 1. Subsequently, it combines these to generate itemsets of length 2 (candidates for frequent itemsets). It calculates their support by scanning the database and eliminates those that do not exceed the threshold. This process continues to generate candidates for frequent itemsets of length 3, and so on[47]. For each frequent itemset, rules are created by considering all possible combinations and subsequently filtering those with confidence greater than or equal to the minimum confidence threshold.

B.5 RelomeDB Supporting Evidence Ranking.

Score	Supporting data
1a	eQTL/caQTL + TF binding + matched TF motif + matched Footprint + chromatin accessibility peak
1b	eQTL/caQTL + TF binding + any motif + Footprint + chromatin accessibility peak
1c	eQTL/caQTL + TF binding + matched TF motif + chromatin accessibility peak
1d	eQTL/caQTL + TF binding + any motif + chromatin accessibility peak
1e	eQTL/caQTL + TF binding + matched TF motif
1f	eQTL/caQTL + TF binding / chromatin accessibility peak
2a	TF binding + matched TF motif + matched Footprint + chromatin accessibility peak
2b	TF binding + any motif + Footprint + chromatin accessibility peak
2c	TF binding + matched TF motif + chromatin accessibility peak
3a	TF binding + any motif + chromatin accessibility peak
3b	TF binding + matched TF motif
4	TF binding + chromatin accessibility peak
5	TF binding or chromatin accessibility peak
6	Motif hit
7	Other

Figure 42. RegulomeDB scoring scheme meaning.

B.6 Filling in the Table of Genes Directly Related to CRC.

- To fill the **"Colon Related Info"** column we refered the output displayed on the **VarElect website**, which provided evidence for the direct association of genes with the phenotype. In some cases, this evidence was sourced from **databases**, while in others, it was extracted from **articles** that VarElect referenced. The task involved citing the database containing the association or, in most cases, extracting relevant sentences from one or more articles to indicate the gene's confirmed or potential role in the disease.
- In cases where the information used to fill this column already provided a general understanding of the gene's function, the **"General Info"** column was not filled. However, when such information was lacking, we used the gene's entry in **GeneCards** (accessible directly from VarElect) to complete this field.
- If **additional data** were found that linked the gene to **cancer in general** but not specifically to colon cancer, this information was added to the "Colon Related" column. For this

purpose, information linking the gene to cancer encompassed not only experimental results but also gene functions related to **cancer hallmarks**.

- Lastly, if VarElect indicated the gene's inclusion in the **cancer census**, the corresponding field was filled with "Yes".

B.7 Management of Gene Aliases to Identify Genes from Our Dataset Present in the Set of Extracted DEGs.

- On one hand, we created a table with the set of extracted DEGs, where their names appeared in two annotations: the external names used by KnowSeq and their translation to Ensembl identifiers [44]. We divided this table into two, the *UP* table and the *DOWN* table, collecting DEGs with positive and negative LFC, respectively.
- On the other hand, we had our 695 genes with three different notations: Ensembl identifiers, external names [45], and symbols (external names used by GeneCards [46]).
- For each of our 695 genes, we assigned a label: "UP" if any of its aliases appeared in the *UP* table, "DOWN" if any of its aliases appeared in the *DOWN* table, and "-" if neither of these conditions was met.

B.8 Louvain Algorithm for Network Community Detection.

Network community detection.

Community detection within a graph involves finding a **partition P of the graph G that generates a modular structure**, which means dividing the network into groups (communities, modules, clusters, etc.) in such a way that there is a higher density of links between nodes within the same group (intra-group) than between nodes in different groups (inter-group).

Louvain algorithm.

In this project, we studied the modular structure of the network by executing the **Louvain algorithm**, one of the most well-known methods for community detection. This algorithm tries to find the network partition into communities that **optimizes the modularity value** (defined as the difference between the number of intra-community links and the expected number of

[44] Obtained using the KnowSeq function *getGenesAnnotation*

[45] Obtained using the KnowSeq function *getGenesAnnotation*

[46] These are the names under which the 695 genes appear in the five tables we created, containing relationships found with CRC or cancer in general.

links in an equivalent random network, considering that the probability of a link between two nodes is proportional to their degrees) [116]. To achieve this, the Louvain method employs a **hierarchical** approach (in each iteration, it creates local-level communities and subsequently aggregates the nodes within the same community into a mega-node) and a **heuristic** approach (it utilizes a neighborhood operator to generate neighboring solutions) [117]. We executed the method using the "modularity" option in the "statistics" tab of Gephi. The chosen resolution parameter, after several tests, was set to 1.7 (higher resolution values result in fewer and larger communities).

C Results and Discussion - Supplementary Material.

C.1 Distribution of Genic SNP Severity Based on the Chromosome.

Figures 43 and 44 illustrate the distribution of the variables "consequence_type_tv" and "Impact", respectively, based on the value of "chr_name".

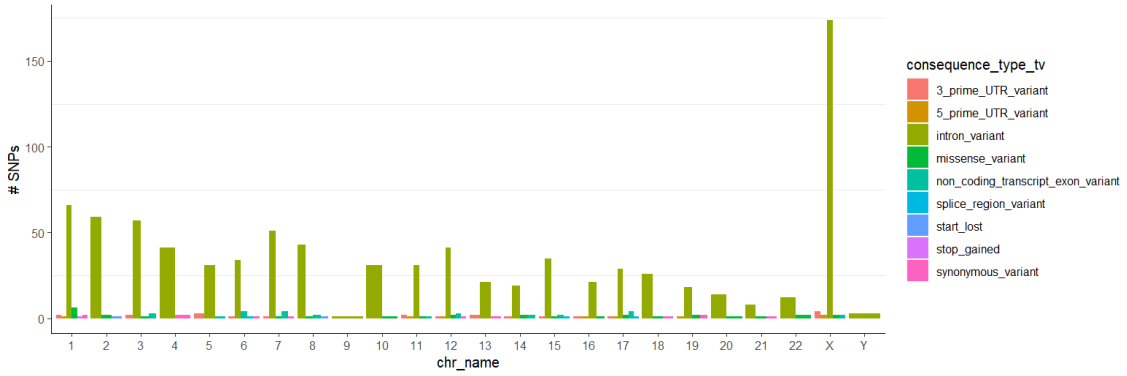


Figure 43. Genic SNP Ensembl table: Distribution of the variable "consequence_type_tv" based on the variable "chr_name"

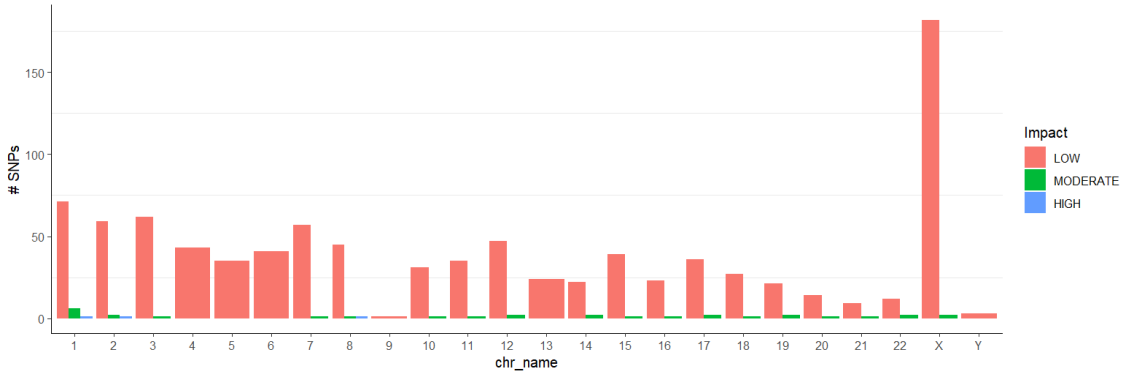


Figure 44. Genic SNP Ensembl table: Distribution of the variable "Impact" based on the variable "chr_name"

Figure 44 informs us that **chromosome 1** accumulates more SNPs with **moderate impact** than chromosome X. Specifically, this moderate impact on chromosome 1 comes from SNPs with **"missense_variant"** as the transcript consequence type (figure 43). A missense variant is a sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved [44].

Furthermore, in figure 44, we can also observe that the **3 SNPs with high impact** are located on chromosomes 1, 2, and 8. From figure 43, we can infer that the "consequence_type_tv" value for these SNPs is **"stop_gained"** for chromosome 1 and **"start_lost"** for chromosomes 2 and 8. These are nucleotide variants that affect stop and start codons. Naturally, variations in start and stop codons are associated with a high impact, as they can lead to the production of completely different proteins, with the consequent effect on protein function [118].

C.2 Protein-Protein Interaction Network Derived from the Total Set of Affected Genes.

C.2.1 Topological and Centrality Analysis.

- The graph, being undirected and having **850 nodes**, has (850*849)/2 = 360,825 possible edges, of which only 2,627 are present, resulting in an **edge density** ($\frac{\text{number of edges}}{\text{number of possible edges}}$) of **0.00728**.
- The network **diameter is 7**, indicating that 7 is the maximum number of edges needed to connect any two nodes in the network. However, on average, two nodes are connected by **3.24** edges (**average path length**).

- The **average degree** (average number of neighbors for a node in the network) is **6.184**, although nodes with higher degrees have many more neighbors (UBC has 232, HNF4A has 161, and FN1 has 111). The network appears to be scale-free, and these nodes act as hubs.
- The **average clustering coefficient is 0.035**, indicating that, on average, 3.5% of the pairs of neighbors of a node in the network are also neighbors of each other (only 3.5% of possible triangles are fulfilled). This is a low value, suggesting a lack of local modular structure in the network.
- The nodes with the **highest eigenvector centrality** in the network (with largest node size in the visualization) are: UBC (1), HNF4A (0.637), and FN1 (0.530). The ranking is repeated for **betweenness centrality**: UBC (0.407), HNF4A (0.210), FN1 (0.113).

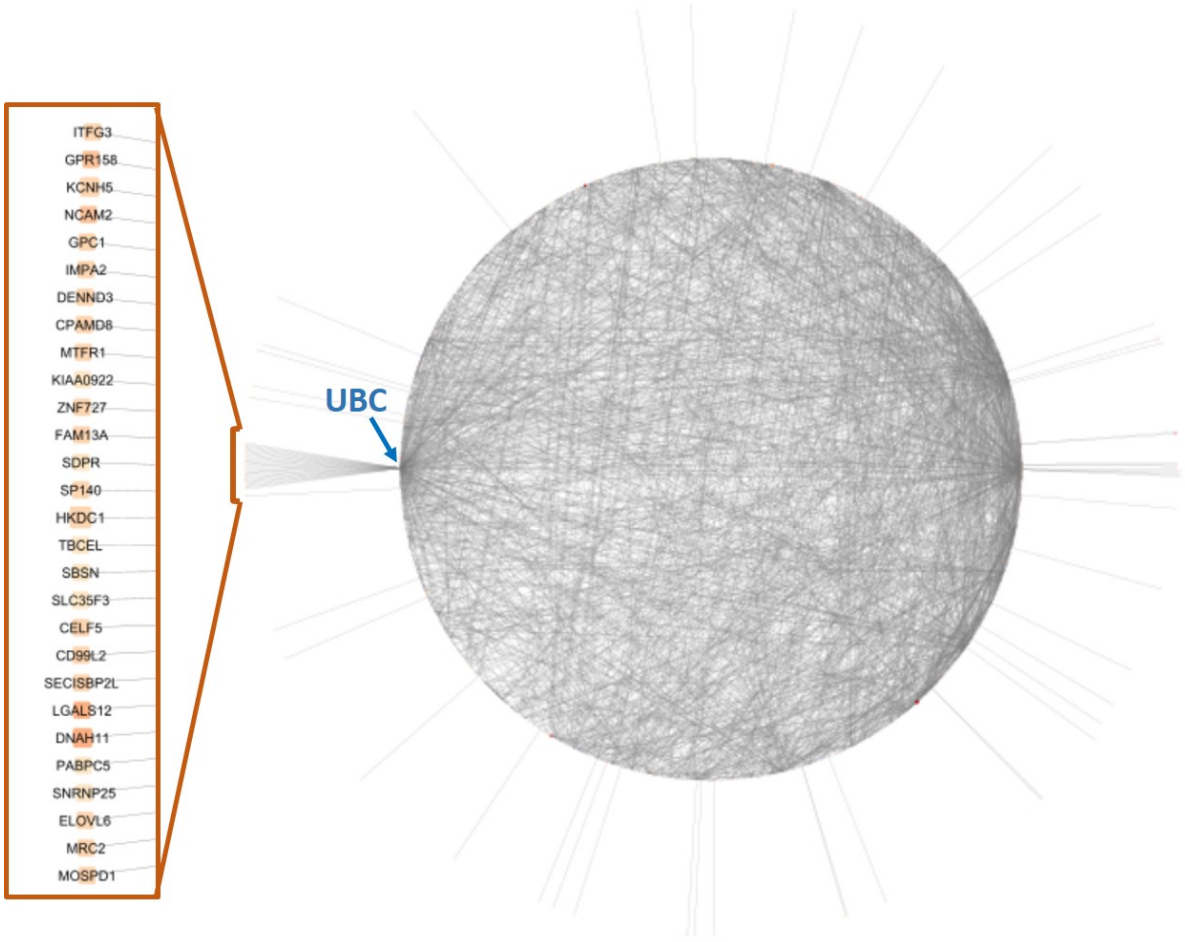


Figure 45. Minimum protein-protein interaction network derived from the total set of affected genes. The circular layout is used to show degree 1 seed nodes conected to UBC. Interactions were obtained using NetworkAnalyst, and visualization was done with Cytoscape.

C.2.2 Key Over-represented Terms in the Nodes.

Table 17. Key over-represented terms in the nodes of the minimum protein-protein interaction network constructed from the total of affected genes. Information extracted using NetworkAnalyst.

Term Source	Over-represented term	Adj p.val
BP	Cell development	2.71E-19
CC	Nucleoplasm	1.13E-25
MF	Enzyme binding	1.18E-31
KEGG	Proteoglycans in cancer	2.45E-13
Reactome	Developmental biology	2.22E-15
Reactome	Signaling NGF (Nerve Growth Factor)	4.67E-15

All of these selected terms have remarkably low adjusted p-values. The term **"cell development"** (BP) and **"developmental biology"** (Reactome) are likely related and may involve the same genes. Furthermore, both may be associated with one of the hallmarks of cancer, "Resistance to growth inhibition signaling" (enabling cancer cells to continue proliferating uncontrollably) [119].

The **nucleoplasm** is the gelatinous region of the cell nucleus that surrounds the chromosomes. Various proteins are found within it, playing crucial roles in the regulation of gene expression, maintenance of nuclear structure, and other nuclear functions. Changes in nucleolar size and number have been recognized as known features of many tumor types [120].

The over-represented **KEGG term** is directly linked to cancer. The KEGG database itself states that many proteoglycans in the tumor microenvironment contribute to the biology of various types of cancer, including proliferation, adhesion, angiogenesis [47], and metastasis, influencing tumor progression.

Nerve growth factor (NGF) -another neurological term- is among the most crucial biologically active molecules in the nervous system. It holds clinical significance in controlling the growth, development, differentiation, and survival of nerve cells, as well as in the regeneration and repair of injured nerves. The expression of NGF has been detected in various tumors [121].

[47] Angiogenesis means blood vessel formation. Tumor angiogenesis is the growth of new blood vessels that tumors need to grow, constituing one of the hallmarks of cancer [119].

C.3 Bicluster-Specific Clinical Data.

C.3.1 Bicluster 16.12 Clinical Data.

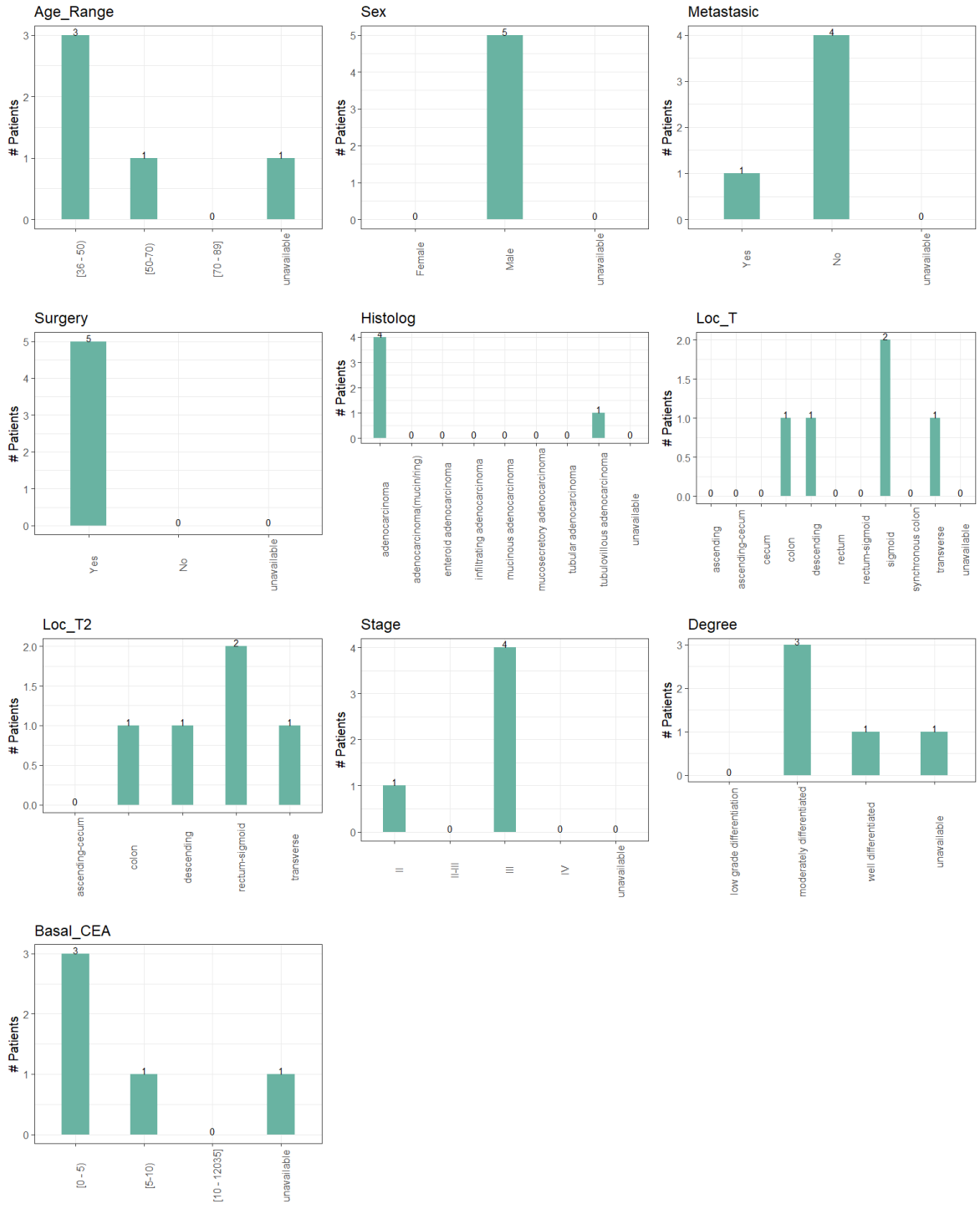


Figure 46. Bar charts of several clinical attributes in bicluster 16.12: "Age_Range", "Sex", "Metastasis", "Surgery", "Histog", "Loc_T", Loc_T2", "Stage", "Degree", "Basal_CEA".

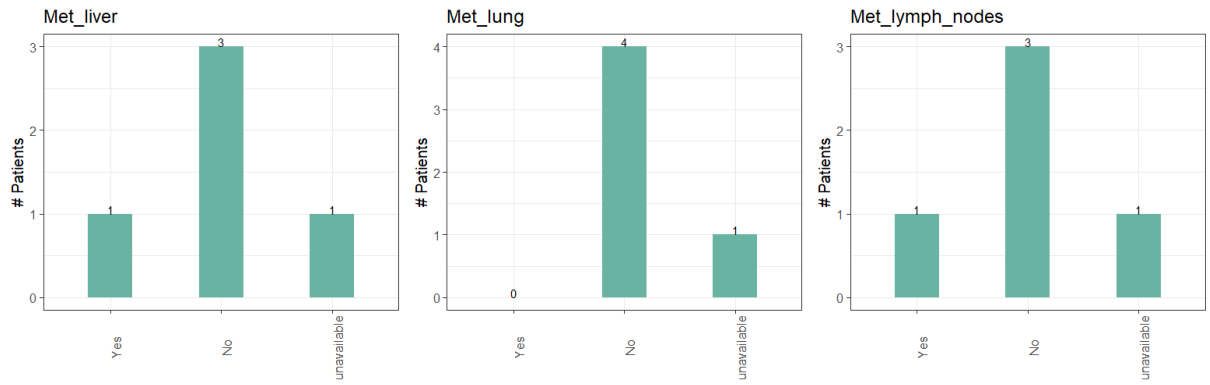


Figure 47. Bar charts of several clinical attributes in bicluster 16.12: "Met_liver","Met_lung", "Met_lymph_nodes".

C.3.2 Bicluster 17.8 Clinical Data.

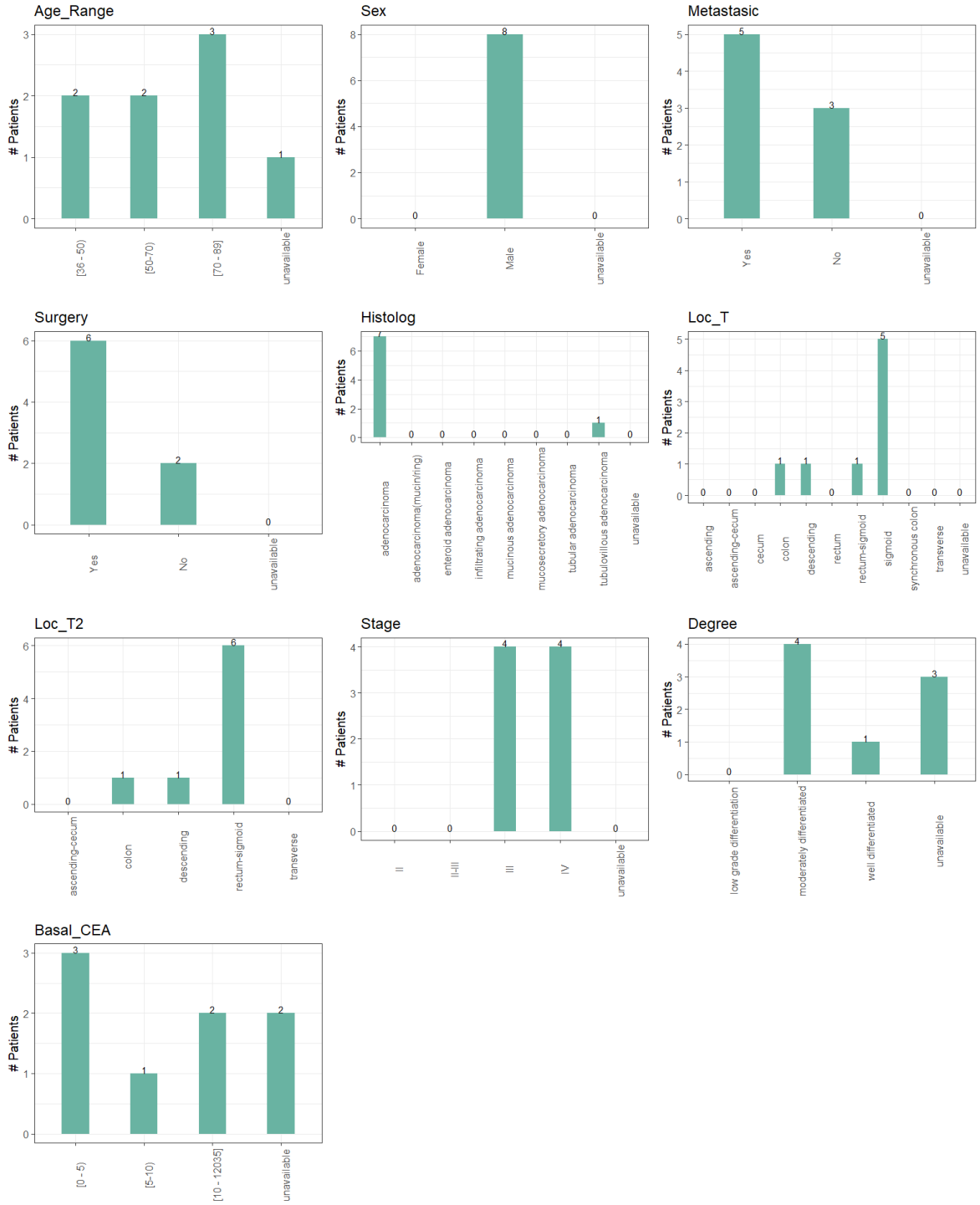


Figure 48. Bar charts of several clinical attributes in bicluster 17.8: "Age_Range", "Sex", "Metastasis", "Surgery", "Histog", "Loc_T", Loc_T2", "Stage", "Degree" and "Basal_CEA".

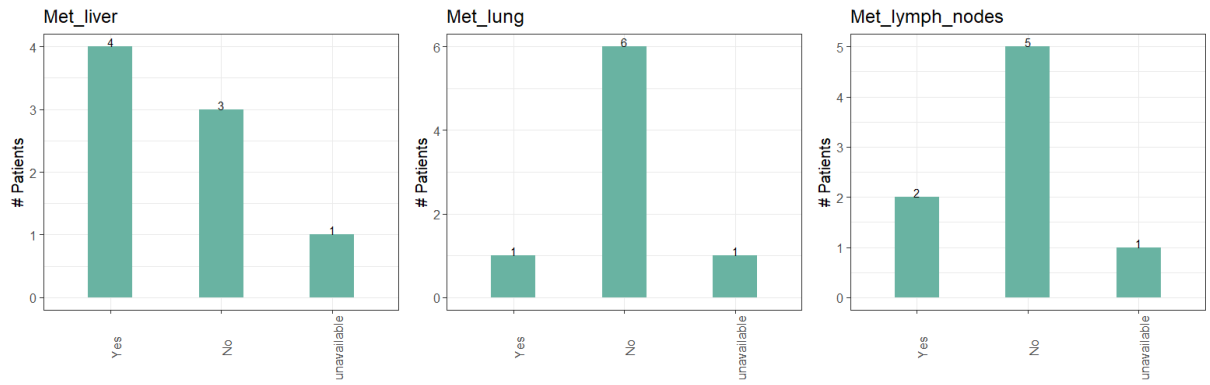


Figure 49. Bar charts of several clinical attributes in bicluster 17.8: "Met_liver", "Met_lung" and "Met_lymph_nodes".

C.4 Interesting Genes.

C.4.1 Interesting Genes Directly Related to CRC.

TNFRSF17 (TNF Receptor Superfamily Member 17) Gene.

- This gene is only included in **biclusters 2.2 and 3.2**.
- It is associated to **a single SNP** (Figure 50).
- This gene's product has been demonstrated to selectively bind to the **tumor necrosis factor** superfamily. It also may transduce signals for **cell survival and proliferation** [122].
- It is directly related to CRC: a study revealed that the **haplotypes of TNFRSF17 SNPs** are associated with **IBD and colon cancer** in a Korean population [123].
- It was identified as a **DEG**, specifically downregulated, in our RNA-Seq data analysis, exhibiting a significant logFC of -3.02.

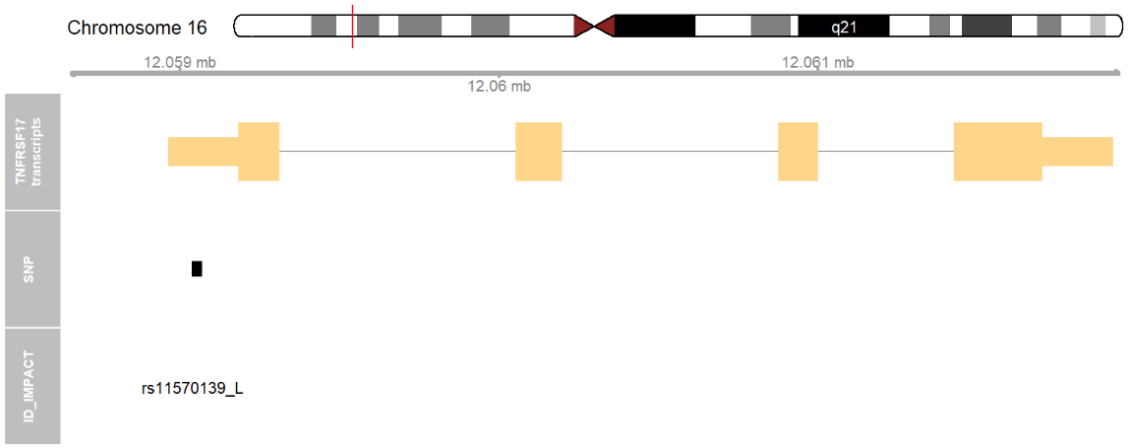


Figure 50. TNFRSF17 gene map. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

GSDMB (Gasdermin B) Gene.

- This gene is included in **7 biclusters** (2.2, 3.2, 10.1, 11.5, 12.1, 14.4 and 15.6).
- It is associated to **5 SNPs** and all of them appear in bicluster 2.2. One of them has **moderate impact**.
- The gasdermins (GSDMs) family is proposed to be pore-forming effector proteins that cause cell membrane permeabilization and **pyroptosis** [48] [125]. Some genes in this family regulate **apoptosis in epithelial cells**, and are linked to cancer [126].
- Diseases associated with GSDMB include **IBD** (according to GeneCards) [126].
- We found a direct association between this gene and CRC: a recent study on the oncogenic role of gasdermins revealed that GSDMB might act as a **negative regulator of cell migration in CRC** [125].

[48] Pyroptosis is a highly inflammatory form of lytic programmed cell death [124].

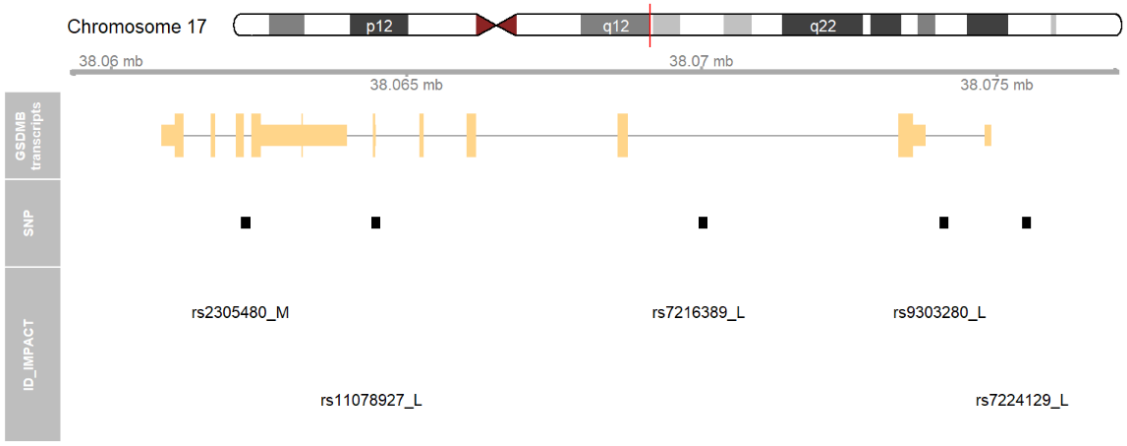


Figure 51. GSDMB gene map of bicluster 2.2. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

DMD (Dystrophin) Gene.

- This gene is present in **12 biclusters** (2.2, 3.2, 8.4, 10.1, 10.4, 11.5, 12.1, 12.9, 13.8, 15.6, 16.12 and 17.7).
- It is associated to **10 SNPs** and 7 of them appear simultaneously in bicluster 2.2 (Figure 52).
- SNP **rs5928111** is present in the majority biclusters.
- This gene's function was **already commented** in Table 12.
- It is **directly related to CRC**, according to ClinVar and DISEASES databases.

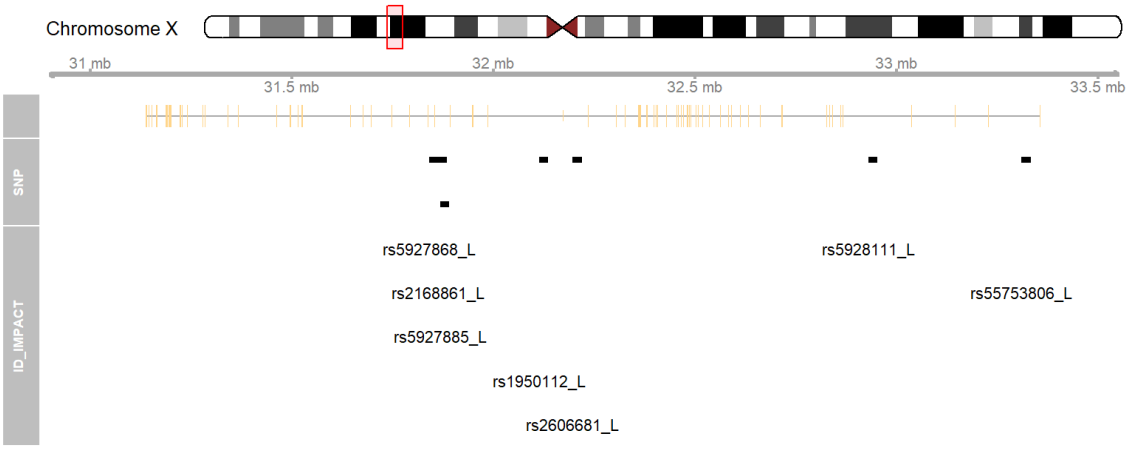


Figure 52. DMD gene map of bicluster 2.2. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

C.4.2 Interesting Genes Indirectly Related to CRC.

PCDH11X (protocadherin 11 X-linked) Gene.

- This gene is included in **10 biclusters** (2.2, 3.2, 8.4, 10.1, 10.4, 11.5, 13.8, 15.6, 16.12 and 17.8).
- It is associated to **3 SNPs** and all of them appear in bicluster 11.5 (Figure 53).
- This gene's function was **already commented** in table 12.
- According to VarElect, it is **indirectly related to CRC through 5 genes**: TP53, CDH1, H19, XIST and CTNNB1.
- Additionally, the presence of other genes encoding **cadherins or protocadherins** in the overall set of affected genes is noteworthy. We identified: CDH9, CDH13, PCDH9, PCDH10, PCDH15 and PCDH19.

We previously mentioned that the loss of cadherins is associated with cancer invasion and progression. Indeed, one of the potential protein biomarkers for predicting tumor progression in CRC is E-cadherin, encoded by the CDH1 gene [127].

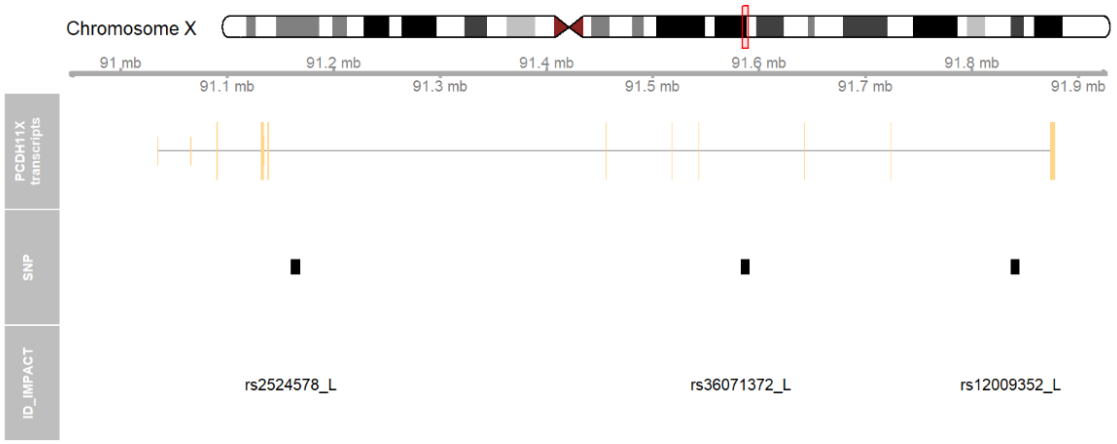


Figure 53. PCDH11X gene map of bicluster 11.5. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

GSTM3 and GSTM5 (Glutathione S-Transferase Mu 3 and Glutathione S-Transferase Mu 5) Genes.

- **GSTM3** appears in **5 biclusters** (2.2, 3.2, 11.5, 12.1, and 16.12), and **GSTM5 in 6** (in addition to the biclusters where GSTM3 appears, it also appears in 14.4).
- Each of them is affected by a **single SNP** (Figures 54 and 55).
- The function of these genes **has already been discussed** in Subsection 3.5.2. While GSTM3 is **directly associated with CRC** ("a novel marker for regional lymph node metastasis in colon cancer"), GSTM5's association is **indirect**. Specifically, VarElect reports that GSTM5 is indirectly linked to CRC through the genes GSTM1, SRC, GSTP1, JUN, and PTGS2.

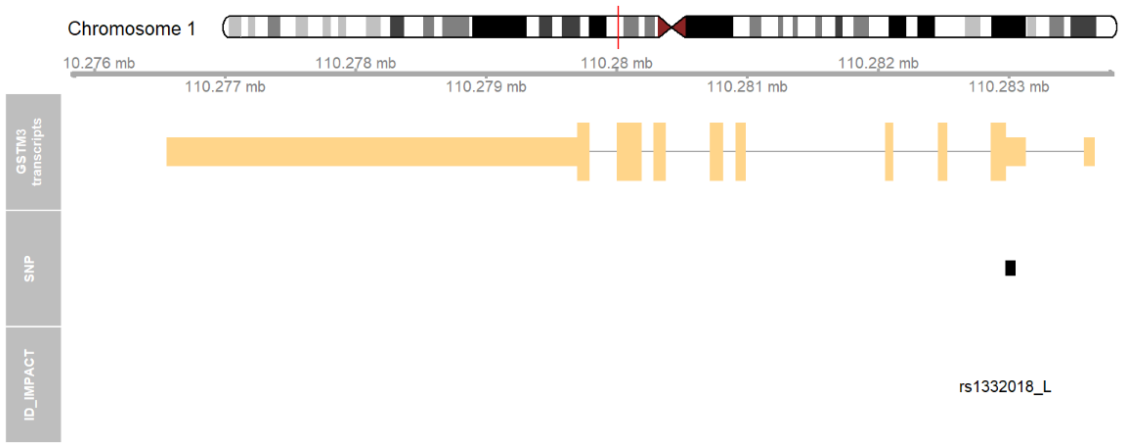


Figure 54. GSTM3 gene map. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

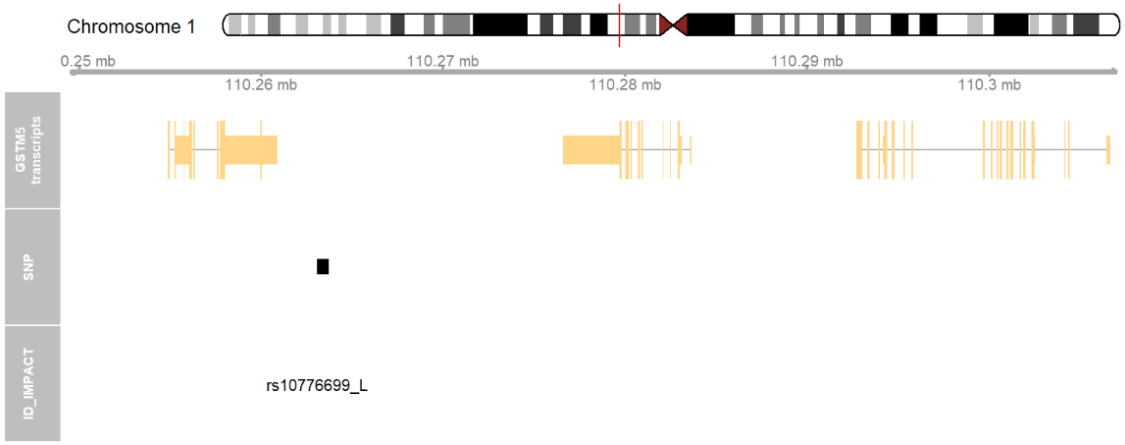


Figure 55. GSTM5 gene map. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

Gene (glutamate ionotropic receptor AMPA type subunit 3) GRIA3.

- This gene appears in **6 biclusters** (2.2, 3.2, 10.4, 12.9, 17.7 and 17.8).
- It is associated to **13 SNPs** and 10 of them appear simultaneously in bicluster 3.2. (Figure 56).
- In bicluster 17.8, the MAFs of GRIA3 SNPs are very high, with one of them reaching the value of 1, as previously mentioned (Figure 33).

- This gene's function was **already commented** in Table 14.
- According to VarElect, it is **indirectly related to CRC through 5 genes**: TP53, BRAF, ERBB2, AKT1 and KRAS.
- It was identified as a **DEG**, specifically underregulated, in our RNA-Seq data analysis, exhibiting a logFC of -2.01.

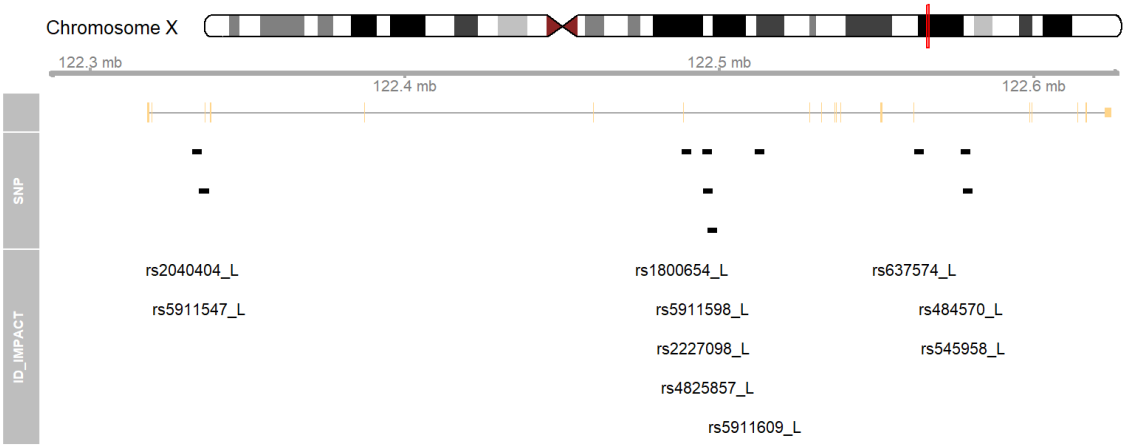


Figure 56. GRIA3 gene map of bicluster 3.2. The top portion displays the relevant chromosome, with the subregion of interest highlighted in red, based on the initial and final genomic positions of the gene. The "Gene transcripts" track presents the combined gene model of the alternative transcripts. At the bottom, the SNPs' locations and labels are plotted along the same genomic coordinate. The letter L, M, or H followed by the SNP identifier denotes the low, moderate, or high impact of the SNP, respectively.

C.5 Brief Selection of Genes that Present Haplotypes in the Genotypic Cluster-Heatmap of Biclusters 2.2 and 8.4.

Table 18. Some of the genes that seem to present haplotypes in the genotypic cluster-heatmap of biclusters 2.2 and 8.4.

ENSEMBL	SYMBOL
ENSG00000077713	SLC25A43
ENSG00000101076	HNF4A
ENSG00000102362	SYTL4
ENSG00000140015	KCNH5
ENSG00000144619	CNTN4
ENSG00000147255	IGSF1
ENSG00000151067	CACNA1C
ENSG00000154310	TNIK
ENSG00000155966	AFF2
ENSG00000168702	LRP1B
ENSG00000182674	KCNB2
ENSG00000184005	ST6GALNAC3
ENSG00000246022	ALDH1L1-AS2
ENSG00000259420	ENSG00000259420