Examining interactions among SNPs that can explain the prognostic variability in colorectal cancer

By

Aaron Albert Curtis

A thesis submitted to School of Graduate Studies in partial

fulfillment of the requirements for the degree of

Master of Science in Medicine (Human Genetics and Genomics)

Division of BioMedical Sciences, Faculty of Medicine

Memorial University of Newfoundland

June 2023

St. John's Newfoundland and Labrador

Abstract

Background: Colorectal cancer is a significant medical burden worldwide and in Newfoundland and Labrador. Examining the relationships of SNP interactions with survival outcomes can help identify new prognostic markers for this disease.

Objectives: To examine associations between colorectal cancer survival outcomes and interactions of SNPs from MMP family and VEGF interactome genes using data-reduction methods.

Methods: Two data-reduction software programs, Cox-MDR and GMDR 0.9, were applied to the data of patients from the Newfoundland Familial Colorectal Cancer Registry. Eight datasets were investigated: one for the MMP gene SNPs (201 SNPs), and seven for the VEGF interaction networks (total 1,517 SNPs). Significance of interaction models was assessed using permutation testing. Associations between significant interaction models and clinical outcomes were confirmed using multivariable regression methods.

Results: For the MMP dataset two multi-SNP models and one single-SNP model were identified, while fifteen novel multi-SNP models and thirteen single-SNP models were identified for the VEGF interaction network datasets. All but one of these models were able to distinguish patients based on their outcome risk in multivariable regression models (p-value range: 0.03 – 2.2E-9).

Conclusion: This research demonstrated that novel genetic interactions associated with outcome risk in colorectal cancer can be found using data-reduction methods. This proves the utility of these methods in prognostic research.

General summary

As a common disease, colorectal cancer impacts individuals, populations, and healthcare systems. Despite its importance, however, little research has been done which examines the combined effects of genetic features in patient survival outcomes. This field is under-studied due to technological and statistical limitations. The computational Multifactor Dimensionality Reduction (MDR) methods were developed specifically to address these issues.

I studied and compared two MDR-based computer programs. Using these programs, I examined approximately 90 million possible combinations of genetic features in a colorectal cancer patient set.

My research found that the MDR programs I used produced different results. My research also identified several previously unknown combinations of genetic features. These are novel potential prognostic markers in colorectal cancer. This is the largest study ever conducted of this type in colorectal cancer. Further studies and progress in this exciting research field should be encouraged.

Acknowledgements

I would like to thank my supervisor, Dr. Sevtap Savas for her guidance, for teaching me how to attain a high level of work quality, for providing me with many opportunities for growth, and for her encouragement for me to develop my skills outside of my program. I'd like to thank the co-authors on my paper, Dr. Sevtap Savas, Dr. Yildiz Yilmaz, Dr. Patrick Parfrey, Megan Carey, and, in particular, my friend and office-mate Dr. Yajun Yu, who, in addition to helping me with the analysis for the paper, was a great support throughout my studies. I would like to acknowledge my committee members, Dr. Yildiz Yilmaz and Dr. Lourdes Pena Castillo for their help and feedback over the years, the staff at CHIA whose aid in the use of computing resources was vital for the completion of my larger analyses, and the funders who enabled my research: the Memorial University Seed, Bridge, and Multidisciplinary research funds. Finally, I would like to thank my peers in the program, my family and friends, and in particular my parents, Linda and Albert Curtis, for their support and constant encouragement.

Contents

List of tables

List of figures

List of abbreviations

B	
BA	Balanced Accuracy
C	
CI	Confidence Interval
CIMP	CpG Island Methylator Phenotype
CIN	Chromosomal Instability
CNV	Copy Number Variation
COAD	Colon Adenocarcinoma
CT	Computed Tomography
CVC	Cross-Validation Consistency
D	
DNA	Deoxyribonucleic Acid
DSS	Disease Specific Survival
E	
eQTL	Expression Quantitative Trait Locus
F	
FAP	Familial Adenomatous Polyposis
FIT	Fecal Immunochemical Test
G	
GMDR	Generalized Multifactor Dimensionality Reduction
GWAS	Genome-Wide Association Study
H	
HPS	Hamartomatous Polyposis Syndrome
HR	Hazards Ratio
HREB	Health Research Ethics Board
HWE	Hardy-Weinberg Equilibrium
I	
indel	Insertion/Deletion
L	
LD	Linkage Disequilibrium
M	
MAF	Minor Allele Frequency

MAP	MUTYH-Associated Polyposis
MDR	Multifactor Dimensionality Reduction
MFS	Metastasis-Free Survival
MMP	Matrix Metalloproteinase
MMR	Mismatch Repair
MSI	Microsatellite Instability
MSI-H	Microsatellite Instability-High
MSI-L	Microsatellite Instability-Low
MSS	Microsatellite Stable
N	
NFCCR	Newfoundland Familial Colorectal Cancer Registry
NL	Newfoundland and Labrador
O	
OR	Odds Ratio
OS	Overall Survival
R	
RAM	Random Access Memory
READ	Rectal Adenocarcinoma
RFS	Recurrence-Free Survival
RNA	Ribonucleic Acid
S	
SNP	Single Nucleotide Polymorphism
SPS	Serrated Polyposis Syndrome
T	
TBA	Testing Balanced Accuracy
TCGA	The Cancer Genome Atlas
TNM	Tumor, Node, Metastasis
V	
VEGF	Vascular Endothelial Growth Factor
VEGFR	Vascular Endothelial Growth Factor Receptor

List of appendices

Research outputs and awards

Publications

Aaron Curtis, Yajun Yu, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. Examining SNP-SNP interactions and risk of clinical outcomes in colorectal cancer using multifactor dimensionality reduction based methods. *Frontiers in Genetics.* 13:902217, 2022. doi: 10.3389/fgene.2022.902217.

Conference abstracts

Oral presentations

Aaron Curtis, Yajun Yu, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. Application of Generalized Multifactor Dimensionality Reduction (GMDR) methods to identify potential multi-SNP interactions associated with survival times in colorectal cancer patients (lightning talk). 6th Canadian Cancer Research Conference. Online. November 8 - 11, 2021

Aaron Curtis, Yajun Yu, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. Application of data reduction methods to identify potential interactions between genetic markers in colorectal cancer outcomes (oral presentation). 22nd Annual Aldrich Interdisciplinary Graduate Research Conference. Online. August 16 - 25, 2021

Aaron Curtis, Yajun Yu, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. Examining the multi-loci predictors of disease outcomes in colorectal cancer using the Multifactor Dimensionality Reduction (MDR) method (oral presentation). 21st Annual Aldrich Interdisciplinary Graduate Research Conference. St. John's, NL. March 23 - 24, 2019

Poster presentation

Aaron Curtis, Yajun Yu, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. Application of a Generalized Multifactor Dimensionality Reduction (GMDR)-based program to identify SNP-SNP interactions that contribute to mortality risk of colorectal cancer patients (poster presentation). 5th Canadian Cancer Research Conference. Ottawa, ON. November 3 - 5, 2019

Awards

Artwork Competition Winner. *BioMedicine Symposium, Memorial University.* July 2022

The Dr. Roger C. Green Graduate Scholarship in Human Genetics. *Memorial University*. May 2020

Student Engagement Award. *Genetics Seminar Series Winter 2019, Memorial University*. April 2019

Co-Authorship statement

I performed all data management steps, programming, and MDR/permutation analyses, performed the statistical analyses, helped interpret the results, gather and interpret eQTL information, conducted the literature search, drafted the manuscript, and created Fig 1 in the manuscript. I am particularly proud to say I performed all coding required for this project, as well as conducted all MDR runs, performed all logistic and Cox regression analyses, and created all Kaplan-Meier curves, completely independently.

I would like to thank the other authors of my manuscript for contributing to my research and manuscript as follows: Dr. Yajun Yu helped with permutation testing, statistical analyses, gathering eQTL information, and literature search; Megan Carey helped collect and process the patient outcome data; Dr. Pat Parfrey led the Newfoundland Familial Colorectal Cancer Registry, helped collect the clinical and genetic data; Dr. Yildiz E. Yilmaz helped with the statistical/permutation analyses; and my supervisor, Dr. Sevtap Savas conceived the idea, led the study, supervised me/the students and research assistant, helped collect the genetic and outcome data, helped interpret the results and draft the manuscript, and finalized and submitted the manuscript.

Chapter 1: Introduction

1.1 Colorectal cancer

1.1.1 Incidence and mortality rates of colorectal cancer

Colorectal cancer (cancer of the colon or rectum) presents a great burden globally, nationally, and provincially. Globally, colorectal cancer accounts for 10.0% of all cancer cases and 9.2% of all cancer deaths[1]. In Canada, for both men and women, colorectal cancer is responsible for the third-highest number of new cancer cases[2] (13,700 new cases among Canadian men, and 11,100 new cases among Canadian women were projected for 2021[2]). Additionally, approximately 5,300 (11.9%) of male Canadian cancer deaths and 4,300 (10.8%) of female Canadian cancer deaths were expected to be due to colorectal cancer in 2021[2]. Among all Canadian provinces, Newfoundland and Labrador (NL) has both the highest incidence (with an estimated 370 new male cases and 300 new female cases for 2021) and the highest mortality rate (an age-standardized mortality rate of 42.8 for men an 27.9 for women projected for 2021)[2]. Research, as well as new health policies are therefore needed to improve prevention, early diagnosis, treatment, and survivorship care for colorectal cancer in order to help alleviate this disease's negative consequences on individuals and populations. The discovery of new risk factors and prognostic markers is vital to this aim.

1.1.2 Risk factors for colorectal cancer

There are a number of known epidemiological, demographic, and clinical risk factors for colorectal cancer. These include age (chances of cancer development increases

with age), sex (males have higher incidence of colorectal cancer), and diet (e.g., consumption of alcohol, red meat, and processed meat increases risk), in addition to smoking, a sedentary lifestyle, and comorbidities such as obesity, inflammatory bowel disease, and diabetes[3–5]. Ethnicity may also affect the risk for developing colorectal cancer. For example people with African American or Ashkenazi Jewish ancestry have a higher risk of developing colorectal cancer[4,5]. Additionally, those with a history of having colorectal polyps are at a higher risk for developing colorectal cancer[4], as are people who have been exposed to ionizing radiation, such as those who were previously treated with radiotherapy for cancer[5].

In addition to these epidemiological, demographic, and clinical risk factors, genetic factors, including hereditary mutations and germline variations, are known to contribute to colorectal cancer development and increased susceptibility. These genetic factors are discussed in more detail, starting in the next section.

1.1.3 *Hereditary and sporadic colorectal cancers*

There are several known hereditary colorectal cancer conditions. The well-studied examples are Lynch Syndrome, Familial Adenomatous Polyposis (FAP), MUTYH Associated Polyposis (MAP), Hamartomatous Polyposis Syndrome (HPS), and Serrated Polyposis Syndrome (SPS)[6].

Lynch Syndrome is caused by numerous variants affecting the DNA Mismatch Repair (MMR) genes, predominantly *MLH1* and *MSH2*[7], which repair DNA errors and damage. Lynch Syndrome patients are at risk of several types of cancer in addition to

colorectal cancer, such as endometrial cancer in women who carry Lynch Syndrome mutations[8]. Approximately 3% of colorectal cancers are caused by Lynch Syndrome[7]. The next most frequent hereditary colorectal cancer after Lynch Syndrome, constituting roughly 1% of colorectal cancers, is FAP, which is caused by mutations that inactivate the *APC* gene[9,10]. Most FAP cases are due to truncating *APC* mutations. FAP is typified by the presence of many (perhaps thousands) of colorectal polyps, and almost always leads to colorectal cancer if untreated (i.e., high penetrant *APC* mutations)[10]. MAP affects less than 1% of colorectal cancer patients[9]. It is similar to FAP, but it is caused by inherited mutations in the *MUTYH* gene[11]. HPS has multiple rare sub-syndromes and may involve mutations in *SMAD4*, *BMPR1A*, *PTEN*, and/or *SKT11* genes[6]. Finally, SPS is characterized by serrated polyps[12]. Several genes have been identified which are associated with SPS (particularly *RNF43*[12]), but known germline mutations explain a very low percentage of cases[12]. These hereditary syndromes have variable degrees of penetrance, but are characterised by early age of onset (and often a positive family history). Since the causative genes are known in these hereditary syndromes, molecular diagnosis and genetic testing are possible[6].

While a small portion of colorectal cancers are caused by known inherited germline mutations, the majority of colorectal cancers are sporadic (~75%[13]), occurring in individuals without an apparent or significant familial history. Intense research to identify the genetic contributors of sporadic colorectal cancers is ongoing[14–16]. In most cases, it is assumed that the sporadic cases are caused by the combined effects of low-penetrant susceptibility alleles and life-style factors[17–19]. Since the causative genes are not

known or have low-penetrance, sporadic cancer patients and their families cannot yet be offered molecular diagnoses or genetic testing, but future studies and clinical applications using polygenic scores may change this[18].

1.1.4 Biological mechanisms involved in initiation, development, and progression of colorectal tumors

In addition to the genes that are responsible for the hereditary cancers, many genes are known to function in the initiation, development, or progression of colorectal tumors[20]. For example, in The Cancer Genome Atlas (TCGA[21]) colon cancer (COAD) and rectal cancer (READ) datasets, *APC*, *TP53*, *KRAS*, *MUC16*, *PIK3CA*, *FAT4*, *LRP1B*, *CSMD3*, *FAT3*, and *FBXW7* constitute the most frequently mutated genes in primary tumors (**Figure 1.1**). Many of these and other mutated genes (including epigenetic mutations) drive colorectal tumorigenesis, leading to the development of carcinomas from adenomas. Depending on the molecular alterations, colorectal cancer is divided into multiple molecular subtypes, including CIN (chromosomal instability pathway), MSI (microsatellite instability pathway), and CIMP (CpG island methylator phenotype)[22]. Molecular subtypes have implications for patient outcomes. For example, colorectal cancer patients with MSI-High (MSI-H) tumors often have better prognosis, even though they do not respond to the primary chemotherapeutic agent (5-fluorouracil)[23].

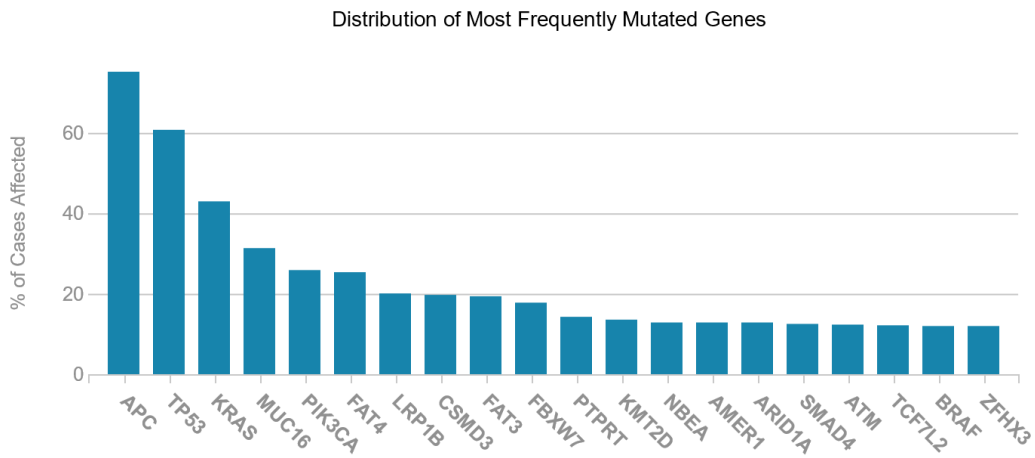


Figure 1.1: The top 20 genes that are mutated in COAD and READ tumors in the TCGA dataset. Generated using the GDC data portal[24] https://portal.gdc.cancer.gov/ (as of August 2, 2022); Parameters used while creating the figure were: Primary site-colon + rectum + rectosigmoid junction; Program-TCGA; Project-TCGA-COAD + TCGA-READ; Disease Type-adenomas and adenocarcinomas + cystic, mucinous and serous neoplasms + complex epithelial neoplasms + epithelial neoplasms, nos; Sample Type-primary tumor.

1.1.5 Progression and clinical staging of colorectal tumors

A colorectal tumor initiates from one malignant cell that has the ability to undergo unregulated growth and division. Usually, if left unattended, tumors grow over time. As colorectal cancer progresses, tumor cells may invade surrounding, non-tumor tissue, recruit blood vessels, and eventually spread, or "metastasize", to distant locations within the body. This progression is described clinically using the concept of disease stage. Stage information is important clinically as it is one of the strongest indicators of colorectal cancer prognosis.

There are multiple staging systems for colorectal cancer. The most prominently used staging system, the TNM (Tumor, Node, Metastasis) system, gives the following breakdown (see **Figure 1.2**): in Stage I, a colorectal cancer tumor initiates and begins to develop and continues to develop through Stage II, growing in size and perhaps becoming increasingly vascularized and/or forming contacts with nearby lymph and blood vessels. In Stage III, cancer cells spread to and invade nearby lymph nodes, and Stage IV is typified by the presence of metastasis, or spread of the cancer to distant organs and tissues[25,26].

The expected survival of patients decreases in each subsequent stage, but once Stage IV has been reached and metastasis has occurred, the chances of successful treatment substantially decreases, and patient survival rates sharply decrease [27]. This dramatic change in survival expectations is a primary motivation for early diagnosis and screening programs, to detect colorectal tumors at early stages of development in order to improve patient survival and disease outcomes.

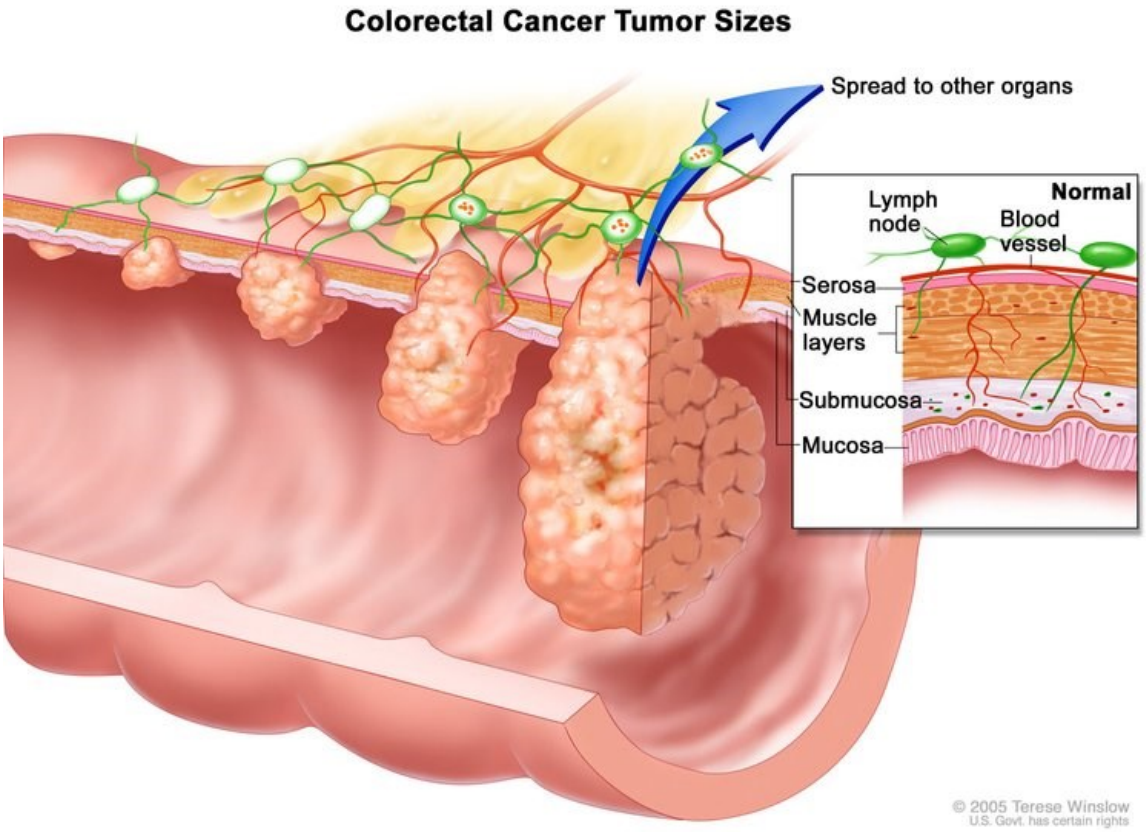


Figure 1.2: Stages of colorectal cancer development. Reproduced with permission from Terese Winslow LLC, https://www.teresewinslow.com/digestion/g5dv4jyf3xcajnghh5xr9lyrm9ghzy. © 2006 Terese Winslow LLC, U.S. Govt. has certain rights.

There are a number of biological pathways involved in tumor growth, progression, and metastasis. Three of these include angiogenesis (growth of new blood vessels), lymphangiogenesis (growth of new lymph vessels), and tissue remodeling. Among the protein families that function in these pathways are the Vascular Endothelial Growth Factor (VEGF) ligands, VEGF receptors (VEGFRs), and Matrix Metalloproteinases (MMPs)[28–31]. As tumor cells can metastasize using blood and lymph vessels as conduits, and tumors need a blood supply to grow, VEGF family genes may be involved in lowered colorectal cancer survival[28,30,31]. MMP genes are involved in the remodelling of the extracellular matrix that holds tumors together, as well as keeps them in place, and thus may also be involved in tumor spread[29]. The established roles of some of these proteins in disease progression has also led to drug development (for example, bevacizumab is an anti-*VEGFA* molecule used in treatment of several cancers, including metastatic colorectal cancers[3,32]). Since they directly influence disease progression, genes functioning in these pathways are excellent candidate genes in prognostic research. This is a primary motivation for the focus on the VEGF and MMP families of genes in my research, in addition to my lab's prior work with some of these genes[33] which allowed me to compare the results of my methodology to the previously used methods.

1.1.6 Screening and diagnosis of colorectal tumors

Clinical screening programs have critical roles in early detection of human cancers. Screening tests for colorectal cancer include faecal occult blood tests, immunochemical tests, stool DNA tests and endoscopy procedures, such as

colonoscopy[3,4]. The most effective screening method currently is colonoscopy, which has the added benefit of allowing the removal of tissue (e.g. biopsy material, potentially cancerous polyps) at the time of screening[4]. Colonoscopy can be followed up with flexible sigmoidoscopy, to screen for cancer in the sigmoidal colon, which has been shown to further reduce mortality[4]. Tests such as stool blood and DNA tests, and CT scans benefit from being less invasive than endoscopic methods, but suffer from a propensity to miss tumors[4].

Screening is particularly important to be offered to patients who have previously had colorectal polyps, colorectal cancer, a family history of colorectal cancer or hereditary colorectal cancer, or other relevant comorbidities[4]. Ultimately diagnosis is performed through biopsy and pathologic examination of suspected cancerous tissue[3]. Since screening methods can help with early diagnosis/detection of colorectal cancer, they allow treatment to begin at an earlier stage of tumor development, and thus dramatically improve the survival of colorectal cancer patients[34]. In Newfoundland and Labrador, eligible residents who are 50 years of age or older can get a FIT kit as part of the Colon Cancer Screening program offered to the population[35].

1.1.7 Treatment of colorectal cancer

Once pathological diagnosis of colon or rectal cancer is made, the next important step is to identify the treatment option that is best for the patient. There are a number of factors which influence clinical treatment decisions in colorectal cancer. These include

disease stage, age of the patient, tumor molecular characteristics, and the affected tissue/tumor location (i.e., colon or rectum).

Treatment for colorectal cancer typically consists of surgery, as well as chemotherapy and/or radiotherapy if needed. Before surgery on the tissue affected by colorectal cancer, neoadjuvant chemotherapy and/or radiotherapy may be performed to reduce the tumor burden. For rectal cancer, typically surgery is performed to remove the entire rectum and much of the surrounding tissue to remove potential lymph node metastasis sites, and this procedure is similar for colon cancer, with sections of the colon being removed[3]. Depending on the disease stage and other indicators (e.g. positive tumor margin), adjuvant chemotherapy may be administered to reduce the chances of recurrence[3].

Common chemotherapy agents include 5-fluorouracil and oxaliplatin. In recent years, targeted therapy, including molecular agents specifically targeting tumor proteins involved in cell growth, division, or metastasis have been approved for use clinically. An example is bevacizumab, an antibody targeting the VEGFA protein[3,32], which is involved in angiogenesis, lymphangiogenesis, and metastasis[36]. Additionally, in recent years immunotherapy has become an effective choice for some patients, particularly those patients who have MSI-high tumors which are not responsive to 5-fluorouracil[37]. As knowledge of the underlying tumor characteristics influences the treatment recommendations made by clinicians, bevacizumab is also considered an agent of precision medicine, where patients are treated based on their individual disease characteristics. These findings emphasize the importance of genetic and genomic

research in cancer and the utility of the molecular information in both drug development and prognosis. Current and future precision medicine-based research, as well as clinical, trials are expected to further increase the number of effective treatment options globally, as well as access to these options, and thus are expected to improve patient survival times. This is particularly important, considering the fact that worldwide cancer incidence rates are expected to increase substantially in the coming decades[38].

1.1.8 Follow up and clinically important survival outcomes in colorectal cancer

Patients diagnosed with colorectal cancer are often followed up for potential disease outcomes by their oncologists or physicians. Usually, the follow up duration is 5-years post-diagnosis[39]. Follow-up examinations and tests (such as blood tests, imaging) are important, as they help monitor disease progression and treatment response/tumor burden, and can detect new recurrences and potential metastases. Generally, recurrence rates for rectal cancer are higher than for colon cancers. Additionally, recurrence and metastasis risks increase with the disease stage at diagnosis[40,41].

As in many cancers from many other anatomical sites, the most commonly analyzed outcome measure in colorectal cancer is Overall Survival (OS), where the outcome event is death from any cause. Other important outcome measures include Disease-Specific Survival (DSS, in which the endpoint is death from colorectal cancer), Recurrence-Free Survival (RFS, in which the endpoint is new tumor recurrence), and Metastasis-Free Survival (MFS, in which the endpoint is new tumor metastasis). By examining the data on end-points and time to patients experiencing the endpoints in a

patient cohort, together with other variables of interest, researchers can examine the relationship and association of variables with the outcomes (i.e., they can identify prognostic markers). There are a number of well-established prognostic markers that are used in the clinic for prognostication. There is also intense research to identify additional markers to refine prognosis and increase prognostic certainty. These are briefly discussed below.

1.1.9 Prognostic markers for colorectal cancer

Prognostic markers are variables which correlate with, and hence can be used to predict, patient outcomes (such as recurrence or death from colorectal cancer). These markers can be demographic characteristics (e.g., age), disease characteristics (e.g., disease stage), tumor characteristics (e.g., MSI), and tumor mutations (e.g., mutations in the *TP53*, *PIK3CA*, *KRAS*, and/or *BRAF* genes)[42].

As mentioned earlier, the disease stage is an important indicator of patient outcomes in colorectal cancer because as the disease stage increases the patients' survival chances decrease[40]. For example, in the United States of America, the 5-year survival rates for stage I, II, and IV patients are 91%, 82%, and 12%, respectively[27]. Survival probabilities also differ among different countries. Overall, in North America, around 60-65% of colorectal cancer patients are alive at the end of their first 5 years post-diagnosis. This number drops drastically in developing countries, however. For example, the 5-year survival rate of colorectal cancer patients in India is around 30%[43].

Age is also an important prognostic marker; as age increases OS chances decrease[44], and biological sex can also affect prognosis – there is a higher mortality rate for men than women[45]. Additionally, patients with MSI-H tumors have better survival times than those with MSS or MSI-L tumors[23]. Finally, mutations in certain genes can predict poorer outcomes as well, sometimes through affecting response to treatment. For example, *KRAS* mutations can be an indicator of chemotherapy response[42] (specifically for cetuximab and panitumumab), which can influence patient survival. *BRAF* mutations can affect immunotherapy response leading to worse prognosis[42]. These examples (that is, the role of tumor MSI-H, *KRAS* and *BRAF* mutation status) once again highlight the importance of biological knowledge in disease management, and its utility in treatment response and prognosis. There is a possibility, however, that our non-tumor DNA harbours additional prognostic markers, such as germline variants.

Similar to tumor somatic mutations, germline genetic variations are exciting candidates in predictive and prognostic research, as they can directly affect the biology of the disease, response to treatment, and/or disease progression. While in many cancers this research area has not yet identified clinically utilized prognostic markers, it is quite promising; several germline variants have been found which influence the effectiveness of colorectal cancer treatment, or predict treatment-induced toxicity, disease progression, or survival. In the next section, I will discuss genetic variations in more detail, particularly Single Nucleotide Polymorphisms (SNPs), which were the focus of my thesis research.

1.2 **Human genome and genetic variations**

The human genome has been explored for many decades, but it was not until the completion of the International Human Genome Project in 2003 that a comprehensive sequence of the human genome was available[46]. While the International Human Genome Project was the first project to sequence the human genome, the project only sequenced the genomes of five individuals, and thus was poorly representative of global genetic diversity. In addition, it suffered from limitations in the sequencing of repetitive regions in the genome, which led to a complete reference human genome becoming available only recently in 2022[47].

Projects subsequent to the Human Genome Project, such as the 1000 Genomes project, sequenced the genomes of far more people (contrary to its name, the 1000 Genomes Project eventually sequenced 2,504 individual genomes[48]) from a broader range of human diversity (26 different populations were included in the 1000 Genomes project, with samples from Africa, East Asia, Europe, South Asia, and the Americas[48]) than did the Human Genome Project. Similar to the Human Genome Project, the research output of the 1000 Genomes project is publicly available, making it a powerful example of open science and the value it can offer. There are several other large-scale projects that sequence human genomes and make the variation information available for public access. For example, the ExAC database includes data from over 60,000 individuals [49,50]. These projects, as well as others, immensely contribute to the understanding of the structure and function of the human genome, genetic features, and the extent of both rare and common

genetic variations, population-based variation differences, and to the conduct of better, more informed genetic research to address health-related problems.

Genome projects have also produced important biological information regarding DNA/RNA coding regions, and non-coding and regulatory regions. Despite the large size of the human genome (approximately 3 billion base-pairs[51]), only approximately 20,000 to 25,000 genes are thought to be protein coding[52]. The typical size of the individual coding regions of a human gene is approximately 1,340 base pairs, but much larger genes exist, the largest being the *TITIN* gene, with a coding region of 80,780 base pairs[46]. While the number of genes is relatively small, many human genes produce multiple proteins through the process of alternative splicing of RNA transcripts, contributing to high biological complexity in *Homo sapiens*[46].

While the total protein coding part of the human genome is small, a significant part of the human genome is transcribed into RNAs. RNAs have important roles in normal development and functioning[46]. They include transfer RNAs which serve in protein translation, ribosomal RNAs which comprise a portion of ribosomes (cellular protein production machinery), small nucleolar RNAs located in the nucleolus (a smaller area within the nucleus of the cell, where the DNA is stored), small nuclear RNAs which are involved in the removal of introns (non-coding segments) from RNA transcripts, microRNAs that are involved in regulation of gene expression, and more[46,53]. RNA species are attracting more and more attention by the scientific community, as interesting new biological and diverse roles emerge for RNAs by the aid of technology and large-scale analysis methods.

While first and foremost, exploration of the protein coding (and, subsequently, the RNA coding) genes was of the highest interest in genetics and biological sciences, recently investigators have also started to realize the importance of the non-coding regions in the genome, such as intergenic and intronic regions[51,54,55]. These regions are known to include regulatory regions and signals and have roles in chromatin structure and accessibility, and hence in regulation of gene expression[48,56]. Overall, the portion of the human genome that has a function is estimated to be up to 80-99%, including regulatory RNA sequences[51,57]. Large scale projects, such as ENCODE[51], have provided invaluable knowledge about the DNA sequences with a regulatory and biological function. This information greatly helps in the interpretation of results obtained by genetic studies and is widely utilized by researchers. Notably, many of the associated signals are located in the non-coding regions of the human genome, suggesting that the low penetrant alleles contribute to human phenotype through alterations of regulatory functions[58,59].

1.2.1 SNPs and other genetic variations

The Human Genome Project and other projects have identified a large number of genetic variations in the human genome. The main types of genetic variants are the Single Nucleotide Polymorphisms (SNPs), Copy Number Variations (CNVs), and short insertions or deletions (indels)[48]. These variations are important for health research – including for susceptibility, treatment response/toxicity, and prognostic studies – as well as in other fields, such as evolutionary biology and population genetics[60].

SNPs are the substitution of a single nucleotide for another. There are more than 84 million SNPs in the human genome[48], making SNPs the most common type of genetic variation in humans. SNPs – as it will be discussed in detail in the next section – have been utilized in health research to find answers to diverse questions, including those of disease susceptibility, treatment response, and prognosis in colorectal cancer.

While SNPs affect a single base-pair, structural variants, on the other hand, are variants which affect segments of DNA larger than a single nucleotide. There are more than 60,000 structural variants in the human genome[48], with a typical individual having over 2,100 structural variants[48]. Examples of structural variations include CNVs, which are differences in the number of copies of a region of DNA between individuals, small insertions/deletions of sequential DNA (indels), and inversions and translocations of DNA regions (segments of DNA which have had their sequences reversed, and segments of DNA which have moved from one genomic location to another, respectively).

Since structural variants include large DNA segments, they are more likely to affect or disrupt genes, and hence may have drastic biological consequences. Examples of such biologically significant structural variants include CNVs that are associated with neurodevelopmental diseases[61,62]. While the vast majority of the studies in cancer examine tumor CNV profiles (e.g., mostly somatic gains or losses of chromosomal regions), a handful of studies have also examined the associations of germline CNVs/indels with prognostic features in colorectal cancer[63–65]. Overall though, SNPs are by far the most studied genetic variation in cancer and other human conditions.

While the number of variants in the human genome is high, relatively few (approximately 500,000[48]) are believed to have functional significance. Some of these are known to alter protein sequence (10,000+ variation sites[48]), but, potentially, most affect gene regulation with no effect on protein coding sequences[48]. This knowledge is one of the reasons why the attention of researchers has recently shifted from protein coding regions to the rest of the genome, particularly to regulatory regions. As mentioned earlier, large-scale collaborative projects, such as ENCODE[51], were instrumental in identifying the regulatory sequences, motifs, and regions along the human genome that are now widely utilized in human health research[51,66].

Finally, it is worth noting that prevalence of specific variants varies from population to population. The most genetically diverse continent is Africa, reflecting the history of human migration[48], but despite differences between populations, most variants show little frequency difference between people of different populations[48]. The difference between two individual genomes is, on average, approximately 1,000 bases[67], and 90 percent of variations are thought to be common (i.e. frequent in the populations or shared by populations)[67]. In health research examining statistical associations, mostly common variants are utilized, as statistical inference examining rare variants individually is quite challenging and requires special approaches[68].

1.2.2 SNPs as susceptibility, treatment response, or prognostic markers in colorectal cancer

As the most common genetic variations, SNPs have been widely investigated to address a variety of research questions in studies related to colorectal cancer. These include susceptibility studies (i.e., examining whether SNPs are associated with or confer risk to the development of colorectal cancer), treatment response and toxicity studies (i.e., examining if SNPs can predict whether a particular treatment is effective in a patient cohort, or examining whether SNPs can predict toxicity induced by treatment), and prognostic studies (i.e., examining whether SNPs can predict a particular patient outcome, such as recurrence). These studies have used candidate variant/candidate gene, candidate pathway, and lately, genome-wide approaches to identify the genetic variations that are associated with colorectal cancer related phenotypes[15,64,69–75]. These studies have created exciting knowledge and have opened new ways to investigate and identify biomarkers.

As an example, there are cases of individual SNPs being associated with the risk of developing colorectal cancer. SNPs at chromosome location 8q24 (most commonly rs6983267), 10p14 (such as, rs10795668), and other loci (e.g. 11q24 and 18q21: rs4939827 and rs7014346, respectively) have been significantly and repeatedly associated with colorectal cancer risk, though the biological mechanisms by which these SNPs associate with colorectal cancer risk is unclear[76]. SNPs in several genes have also been associated with prognosis in colorectal cancer. For example, in genes encoding microRNAs, there have been several SNPs associated with factors related to colorectal

cancer survival, such as metastasis or tumor location[77]. Some research has indicated that SNPs in the *TP53* gene may influence colorectal cancer survival (most prominently, rs1042522)[78], and SNPs in many other genes (e.g. *MTHFR*, *KDR*, *VEGFA, SMAD7*) and non-coding regions have been associated with colorectal cancer clinical survival outcomes as well[79]. For a comprehensive list of studies and their findings, please review the dbCPCO database, a database that catalogues studies examining the associations of genetic variants with treatment response, toxicity, and patient outcomes[69]. In some cases, reported associations have been replicated, increasing the confidence in their potential prognostic value. For example, Negandhi et al. demonstrated that the *MTHFR* Glu429Ala polymorphism was associated with OS and the *ERCC5* His46His polymorphism was associated with disease-free survival two colorectal cancer patient cohorts[71]. Additionally, there is some evidence that the same variants may affect both risk and prognosis. For example, rs4939827 from the *SMAD7* gene, a SNP previously found to be a risk factor for colorectal cancer, was subsequently found to also associate with DSS[80]. Similarly, risk associated variants rs4779584 and rs10795668 were found to correlate with disease outcomes (death and recurrence)[79]. Association of rs4779584 was replicated in several populations[81] and rs10795668 was subsequently associated with OS for colorectal cancer in a different cohort of patients[82]. Since variants that contribute to risk of developing cancer - in at least some cases - can also affect the tumor features, and thus, cancer progression, variants of this type are exciting candidates in prognostic research.

While some studies study a small number of variants, Genome-Wide Association Studies (GWASs) are considered the gold standard in SNP association studies. Published GWAS studies in colorectal cancer survival outcomes are currently limited in number, but GWASs are increasingly being utilized to examine the genome-wide association patterns of SNPs[64,72,73,75,83–86]. In almost all of the published GWASs in colorectal cancer, different sets of variants were identified as being associated with patient outcomes. This may be due to differences in patient cohorts (e.g., disease stage, treatment regimens, ethnicities, or the examination of different outcome measures), or due to the fact that their findings are spurious (i.e. are false-positives) or cohort-specific. To my knowledge, only one recent GWAS has reported association of a genomic region with a disease outcome in two independent patient cohorts[64]. Further studies on this region can reveal its biological relation to disease progression in colorectal cancer, and hence, may open new ways to control this disease.

In summary, investigating genetic variations as susceptibility, treatment response/toxicity, and prognostic markers is quite promising. However, it is important to note that the vast majority of the studies are limited in the sense that they examine the relationship of individual variants with outcome one at a time. Importantly, there also exists the potential for ***interactions*** between SNPs to serve as prognostic markers for colorectal cancer. This potential exists as the complexity of the biology underlying colorectal cancer may possibly result in multiple variants affecting survival in a non-additive manner. To date, very little research has been done in the area of interactions between SNPs serving as prognostic indicators, mostly due to the difficulties associated

with interaction analysis. My thesis research has focused on this challenging but also exciting area of SNP interactions that can predict clinical outcomes in colorectal cancer.

1.3 **SNP interactions**

Single SNP association studies suffer from an obvious limitation: if some variants have a different effect on an outcome when, and only when, in the presence of one or more other variations, this information cannot be ascertained by studying the effects of individual SNPs in isolation (i.e., by examining each SNP's relation to outcome individually), as is done in most of the published studies.

In contrast, SNP interaction studies examine the associations between a response variable of interest (e.g., an outcome variable, like OS) and the genotypes of a single SNP or multiple SNPs simultaneously. Interactions involving the genotypes of a single SNP are referred to as 1-way interactions; interactions between two SNPs are referred to as 2-way interactions, interactions involving three SNPs are referred to as 3-way interactions, and so on. Note that since 1-way analysis includes only one SNP, it is the interactions among SNP's different genotypes that are examined. The associations of interactions to response variables are non-additive (i.e., the effect of an interaction is different from the sum of the effects of each SNP individually), differentiating SNP interactions from the cumulative effect of multiple SNPs.

Biological interactions occur in nature and are sometimes referred to as epistasis[87,88]. Epistasis has roles in evolution of genes[89] and functional compensation[90]. In genetic and health research, epistasis/interactions are also considered to represent the

"missing heritability"[91–93], which is to say that they explain a portion of the genetic basis of a trait which, so far, has not been explained.

SNP interactions are an under-studied area of research, particularly in colorectal cancer. This is due to several challenges, which are inherent to the analysis of interactions (these challenges are discussed in the next section). While there has been little work done studying germline SNP interactions and cancer prognosis[94–98], given the complexity of cancer, it is likely that germline variants, such as SNPs, may interact to influence cancer outcomes, including for colorectal cancer. There are published examples supporting this. For example, Afzal et al. found interactions of polymorphisms in the *DPYD* and *TYMS* genes to be associated with DSS in two cohorts of colorectal cancer patients[94]. The authors caution against biological interpretation of their statistical result[94]. In a follow-up study, Sarac et al. further determined that interaction profiles identified in Afzal et al., for polymorphisms in the *DPYD* gene and *TYMS* gene, were associated with time to death, time-to-relapse and adverse drug reactions in colorectal cancer, based on a comparison of two cohorts[96]. In another study, Pander et al. found an interaction in a dataset of *TYMS* enhancer region and *VEGFA* gene polymorphisms for progression free survival for colorectal cancer patients[95]. This interaction was identified specifically for metastatic patients treated with the drug CAPOX-B[95]. The authors indicated no definitive underlying mechanism for the interaction identified[95]. Hence, an interesting piece of the puzzle, which is the underlying biological basis for the observed interactions, awaits discovery.

There have also been interactions found which associate with colorectal cancer risk. Research by Jung et al. demonstrated an up to 17 times greater risk for colorectal cancer, when stratified by use of oral contraceptives, when individuals had particular variants at both rs1800961 and rs4092465, but not either variant individually[98]. The authors speculate on the biology implicated in their findings; while no mechanism is known, both SNPs are in genes related to colorectal cancer and/or sex hormones (such as estrogen levels, influenced by oral contraceptives). These SNPs are located, respectively, in the *ONECUT2* gene, a known tumor promoter, and the *HNF4A* gene, whose protein product is known to inhibit cell proliferation in colorectal cancer and associated with sex hormone-binding globulin under certain circumstances[98]. This information provides important clues to help further dissect the relationship between interactions and the biology behind their associations with phenotypes. Generally speaking, biological interpretation of interactions is challenging due to the fact that biological information that considers interactions is largely non-existent. Functional studies examining interactions are therefore required in order to further elaborate on disease mechanisms and to develop disease control measures (e.g., new drugs) based on interactions.

In summary, while there are challenges associated with interactions in terms of understanding their biology, identifying them is also a very promising area in health research. Some of these serious challenges associated with interactions are technical in nature and are discussed in the following sections.

1.3.1 Sparse data problem

The sparse data problem is a problem inherent to interaction analysis[99–101]. As the number of variables (e.g., SNPs) being included in an interaction analysis increases, the number of possible combinations of genotypes also increases (for example, if an analysis has one SNP (1-way analysis), there are 3 possible genotypes; if an analysis has two SNPs (2-way analysis), each with 3 possible genotypes, there are 9 possible genotype combinations; and if an analysis has three such SNPs (3-way analysis), there are 27 possible genotype combinations; see **Figure 1.3**.

As the number of possible genotype combinations increases, fewer combination contingency table cells will have sufficient samples available in the dataset to represent that combination. The data, thus, becomes, "sparse": that is, data becomes distributed among a larger number of combined genotype cells, leading to reduced statistical significance and interpretability of the analysis. See **Figure 1.4** for a demonstration of this concept.

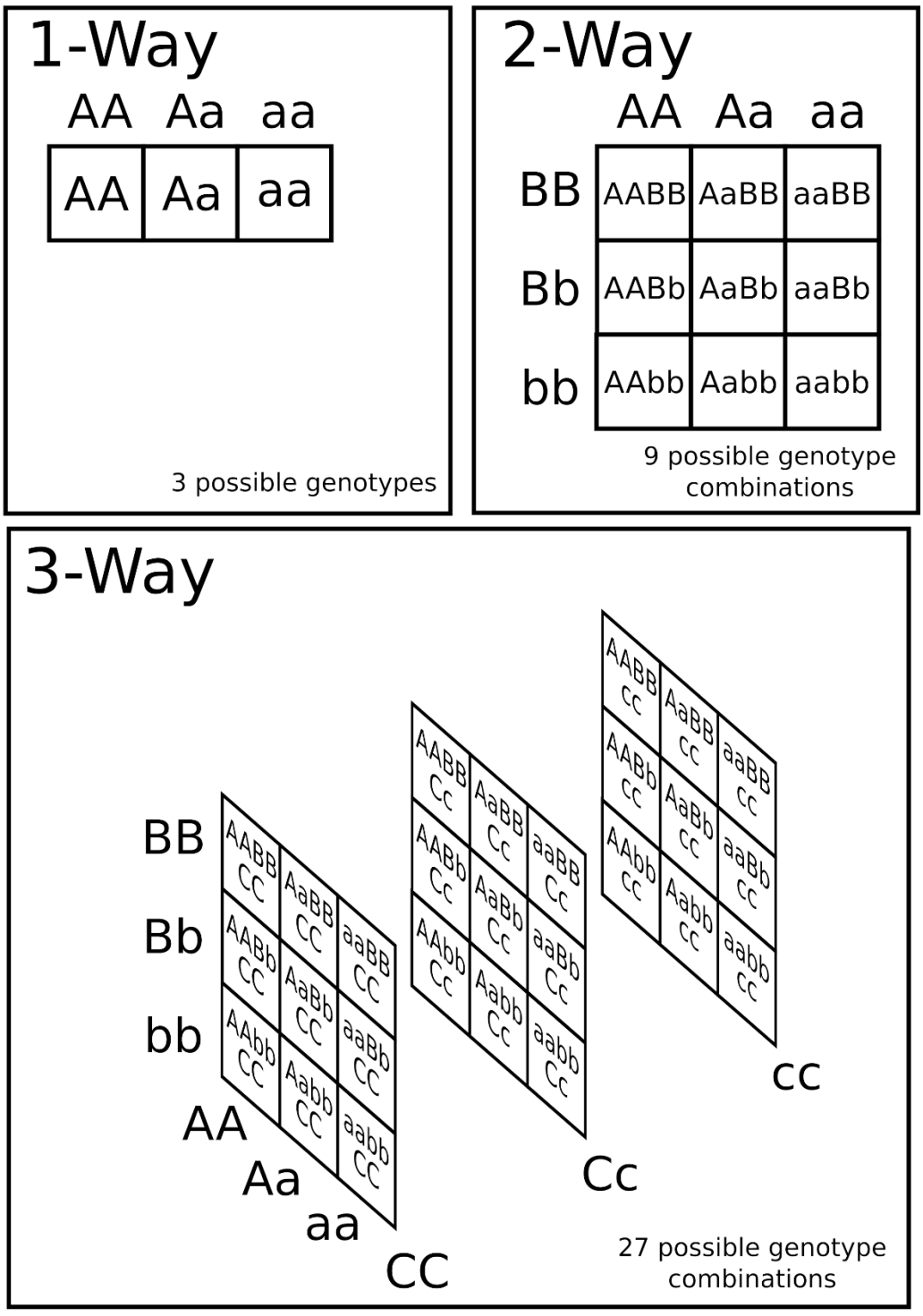


Figure 1.3: Demonstration of all possible genotype combinations of 1, 2, or 3 SNPs in interactions. Each SNP having three possible genotypes (e.g. AA, Aa, or, aa, where A represents the major allele, and a represents the minor allele)

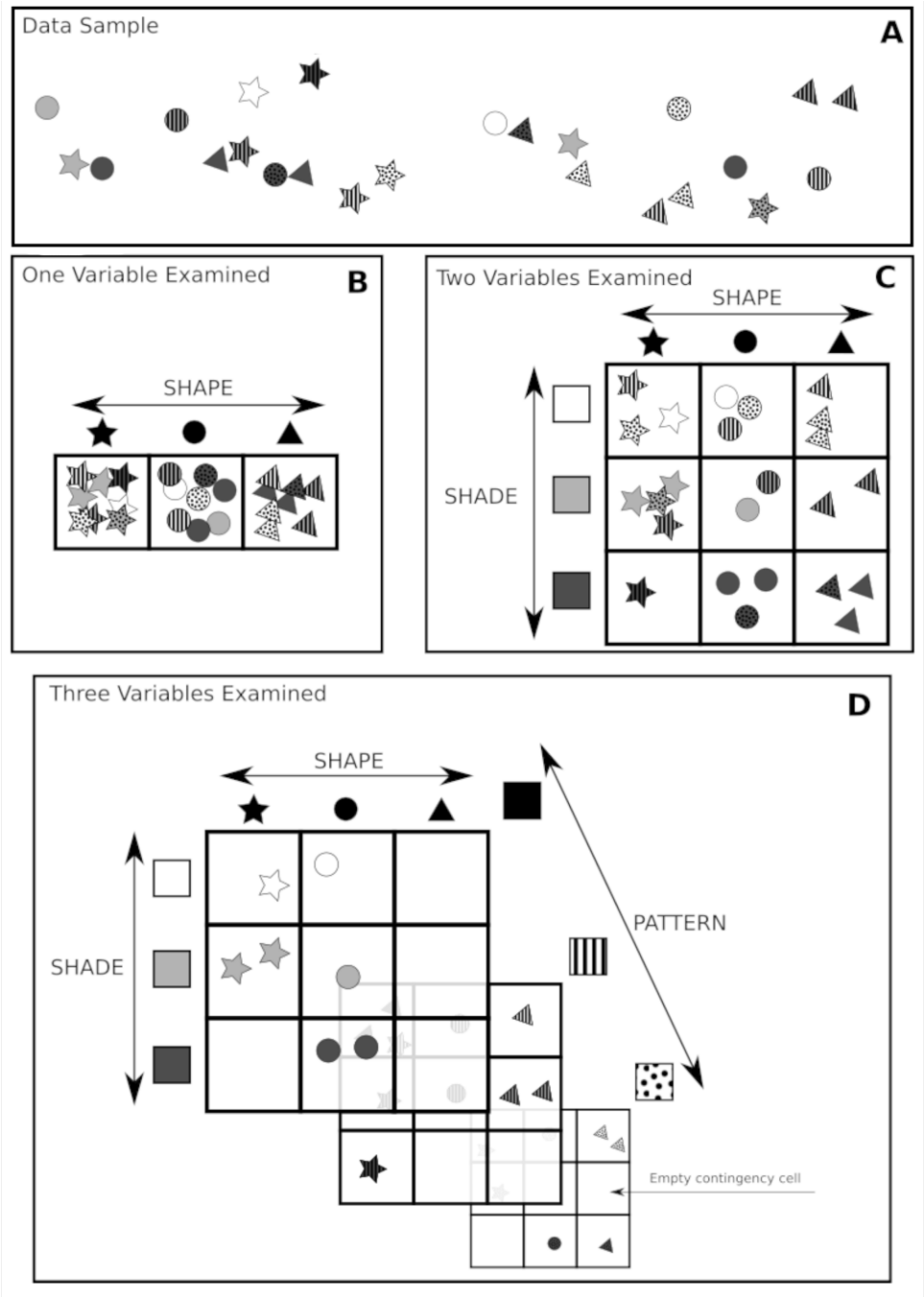


Figure 1.4: Graph demonstrating the increasing sparseness of data as samples become increasingly spread out over an increasingly large number of contingency cells. As the number of variables (e.g. SNPs) increases, an increasing number of possible variable states (e.g. SNP genotype combinations) inevitably have insufficient data. If this analogy is applied to genetic variants, the "Shape", "Shade", and "Pattern" variables can represent different SNPs, and each attribute (e.g., star, gray, striped, etc.) can represent the genotypes of the respective SNP.

1.3.2 Computational complexity

Another challenge in interaction analyses is that as the number of variables (e.g., SNPs) examined in a dataset increases, the number of possible combinations of these variables also increases non-linearly.

As an example, for a dataset with 100 variables/SNPs, the number of 2-way combinations of these variables/SNPs is 4,950, and the number of 3-way combinations is 161,700. If we instead have 200 variables/SNPs, the number of 2-way combinations increases to 19,900, and the number of 3-way combinations increases to 1,313,400. Hence, computationally it can become quite a demanding process to examine interactions in such datasets. Depending on the algorithm being applied, examining large datasets thus may require a large amount of time and/or memory to process. See **Figure 1.5** for a visual demonstration of this growth of complexity.

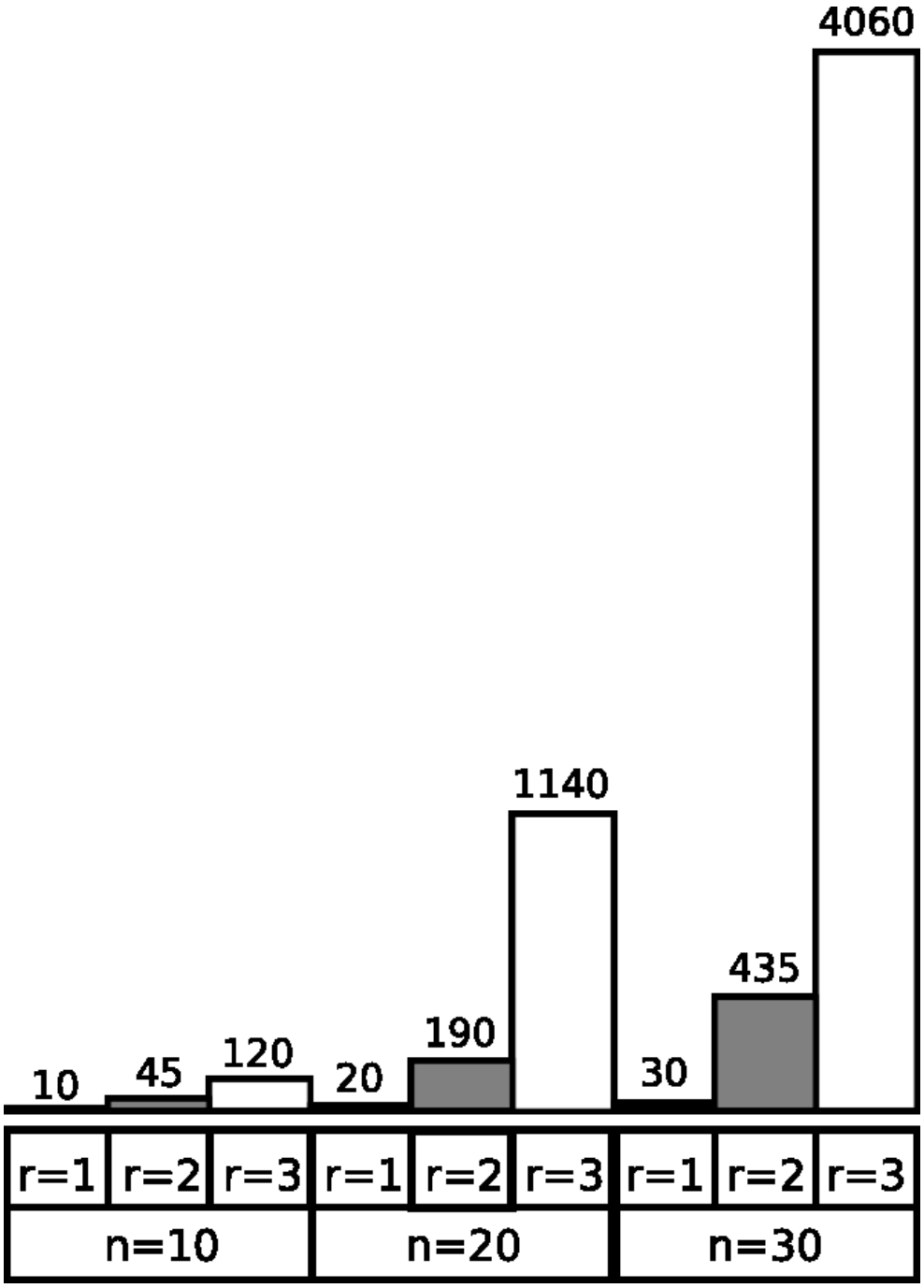


Figure 1.5: Demonstration of the growth in complexity of the combination formula . r is the number of variables constituting a single interaction, n is the number of variables (e.g., SNPs) being studied in total. The number at the top of the bar , and the height of the bar, indicates the number of combinations of size r among the total number of variables n. This number grows non-linearly on both the r and n.

There are several ways to alleviate these problems. The issue of time complexity (the amount of time required to complete the analysis on increasingly large datasets) can often be reduced by the use of multi-processing[102–104], that is, some problems can be solved concurrently in many small, independent pieces before combining these partial results into the final result. Multiple interactions may be examined simultaneously, reducing the time it takes to complete the overall analysis. GMDR 0.9[101], one of the methods I have used in my thesis research, uses multiple parallel-processing threads for this purpose. Space usage may sometimes be reduced by combining partial results into aggregate results, if the details are not needed. I used this approach in my extension of the Cox-MDR[105] code in order to limit the amount of Random Access Memory (RAM) being used, storing only the highest scoring model at a time instead of storing data for all models and picking the best scoring model from them. Some methods also select SNPs, from the total pool of SNPs, to limit the size of the input data to be processed. As I will discuss in the next section, many of these solutions require either strong computational knowledge and resources, or advanced approaches.

1.3.3 Methods for SNP interaction analysis

Traditionally interactions have been explored in longitudinal survival data by the use of methods such as the Cox regression method, and the addition of an interaction term for the variables of interest. While useful for examining one or a few interactions in the same model, this approach is impractical when applied to big data. Therefore, while interaction terms in statistical models can be a useful approach to examine a small

number of interactions, for larger datasets bioinformatics algorithms have been developed to examine interactions while working with or around the limitations of such traditional methods. These algorithms may exhaustively examine all possible interactions, or they may explore a portion of interaction space.

Bioinformatics is a broad research area which encompasses research that utilizes knowledge from both biology and computer science. Bioinformatics research includes a diverse range of topics, and can be used in the discovery of biomarkers, such as prognostic markers. Bioinformatics approaches may also incorporate machine learning, which utilizes computational methods inspired by human learning that find associations in data [106].

Machine learning methods use various procedures to fit models to data such that these models can predict associations in data which is independent from that on which the model was constructed. In a basic form, this can refer to regression methods such as linear regression, but machine learning is often used to model complex data which may not behave in a linear fashion[106]. Machine learning methods are "trained" on data, that is they construct models based on a dataset, and then they are tested, i.e., it is determined whether or not they are capable of predicting an independent dataset[106]. Machine learning methods are often used for feature selection.

Feature selection refers to the process of selecting "features" from a dataset which are of interest[107]. For example, one may want to select a set of SNPs from a dataset which are of interest to cancer prognosis, or one may wish to select a subset of features from a

dataset to reduce the number of features to a specific problem domain or to lower computation burden[107].

Feature selection algorithms may be univariate or multivariate, and may be filter, wrapper, or embedded methods[107]. Filter methods rank computed models, producing a list of the best models as determined by the algorithm, whereas wrapper and embedded methods are more complex at the cost of greater computing resource use[107]. The Multifactor Dimensionality Reduction (MDR) procedure used in my work can be seen as a multivariate filter method[107]; it produces a ranked list of models (returning the "best" one) and operates on multiple features (i.e. multiple SNPs).

In cancer genetics research, several feature selection machine learning algorithms have been employed for biomarker discovery. For example, Chen and Dhahbi were able to find five novel biomarkers which could distinguish gene expression patterns between lung adenocarcinoma and lung squamous cell carcinoma using five feature selection algorithms[108]; Feng et al. applied feature selection to transcriptomic, clinical, and molecular data from colorectal cancer patients to predict survival and progression[109]; and Al-Rajab et al. compared several feature selection methods on the task of colorectal cancer diagnosis from gene expression data[110].

Exhaustive search methods, such as MDR and FastANOVA[99,111,112] work by looking at all possible combinations of variables in a dataset. If every possible interaction in the dataset is to be considered for its potential to have a significant statistical effect, an exhaustive search of the interaction space is required. These methods have a clear downside, though, if all combinations are to be considered, then the computational

burden of the problem will be high, as some amount of work must be done for each possible combination. While MDR, for example, reduces computational resource requirements somewhat by reducing the combination of genotypes, for each combination of SNPs, to a single dimension, aiding in model comparison, the software must still perform computation for every possible combination. As such, there are other methods which, while potentially less robust, have been developed to find interactions without exhaustively searching all possible combinations of variables. Some of these methods are described in the remainder of this section.

The random forests method works by using random subsampling of patients and by constructing trees for many of these subsamples based on using SNPs as classifiers, with these SNPs being selected from a random subset of the total dataset[113]. These collections of trees, or "forests" are then used to determine likely interactions[113,114]. While the random forests method cannot detect interactions which only contain SNPs which do not have main effects[111], they have also been used to screen SNPs, to reduce the total number of SNPs before performing other statistical interaction analyses[113,115].

Use of Artificial Neural Networks, inspired by biological neural networks, is another method to examine interactions without resorting to a brute-force approach. These networks are trained using known associations and that training is applied to finding novel associations[111,116]. Various Artificial Neural Network methods have been applied to the problem of interactions with mixed success[114].

Ant Colony Optimization, like Artificial Neural Networks, is a biology inspired algorithm, finding SNP interaction associations in a dataset in a way that is similar to

how biological ant colonies find shortest paths to food sources[117]. This method offers high power and a low false positive rate, but suffers from the same issue of scaling with interaction order as regression methods do[111].

As discussed earlier, regression models may be used to study interactions by incorporating interaction representation into traditional statistical models[111], yet they suffer from a need for large samples and model-dependency[114]. In addition to frequentist approaches, interactions can be incorporated into Bayesian statistics as well. Multiple Bayesian-based methods exist, such as BEAM[118], FEPI-MB[119], and bNEAT[120] but, similar to the random forests method, interactions found using these methods must also contain main effects, limiting these methods from finding interactions between variants which do not have individual effect[111].

The methods listed above, and other methods, show that there are many methods which have been developed to examine the interactions, each with its own limitations and strengths. I chose to work with Multifactor Dimensionality Reduction (MDR)-based methods for my work for this thesis.

1.4 **Multifactor Dimensionality Reduction (MDR) method**

1.4.1 The MDR algorithm

MDR is a data reduction method first developed by Ritchie et al.[99] to study interaction effects while mitigating the sparse data problem. It accomplishes this by pooling available data into two categories, high-risk and low-risk, for a particular measure. This measure may be susceptibility to a disease, as the term “risk” often implies

in genetics research, but may also be a survival outcome measure, where high-risk and low-risk refer to the "risk" of poor survival outcomes. MDR takes the relatively high-dimensional data inherent to combinations of SNP genotypes (e.g., genotypes of multiple loci) and *reduces* it to a single dimension—high risk genotypes vs. low risk genotypes and an associated score. MDR is designed to do this for every possible combination of explanatory variables in a dataset for a given order (or up to and including that order) of an interaction (e.g., each 1, 2, and 3-variable combination in the dataset). MDR accomplishes this data reduction via a scoring system, which rates a particular combination of SNPs, and its associated genotype values, for its ability to distinguish patients who are at high-risk for a particular disease outcome from patients who are at low-risk. MDR is an example of a machine learning algorithm, as it "learns" by training a model on a dataset, and then uses this model to predict associations in independent datasets.

For the MDR algorithm, a cross-validation approach is implemented in order to evaluate the models constructed by the algorithm. In cross-validation, data samples are split into a number of partitions. All but one of these partitions will be designated as a training set, and the remaining partition will be designated the testing set. Subsequent steps are repeated as many times as there are partitions, using a different partition as the testing set each time. The rationale behind cross-validation is that by testing each model on an independent set of testing data (i.e., which was not part of the training data used in the model's construction), one can determine whether the model is likely to be a true

model, as a true model should be able to be replicable in datasets other than the one used to develop the model[99].

While examining SNP interactions, using the data from the training set, a model is constructed for a particular combination of SNPs via a simple ratio of the number of case samples (i.e., high-risk group) with a particular genotype combination to the number of control samples (i.e., low risk group) with the remaining genotype combination (see **Figure 1.6**). This procedure is repeated for every combination of SNPs in the dataset (thus exhaustively searching the interaction space). Upon generation of MDR models, models are then evaluated on their ability to distinguish the high-risk from low-risk patients in the testing set. Thus, a high-scoring model is one which can distinguish patients independently of the data that was used to generate it. This is achieved by calculating the prediction error for the model.

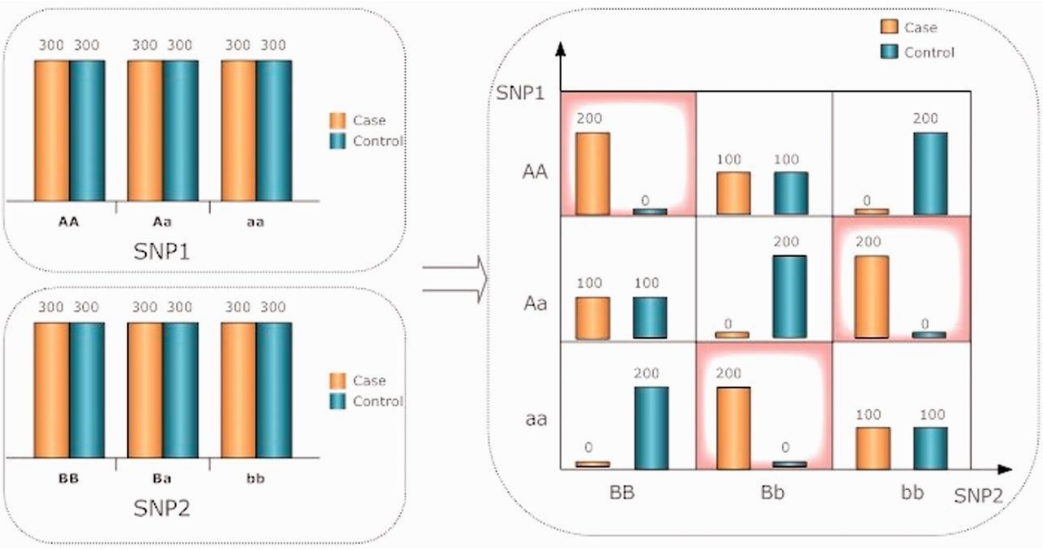


Figure 1.6: This figure demonstrates the selection of "high risk" genotype combinations by comparing ratios of cases to controls for a particular response (i.e., if there are more cases than controls, for a given response variable, with a particular genotype combination, that combination is deemed "high risk", otherwise it is "low-risk"). Figure reproduced from Li et al. (2016)[87]. Reproduced with permission of the publisher, Oxford University Press.

In the cross-validation procedure, the construction and scoring of models is repeated such that each of the data partitions constitutes the testing set exactly once[99]. At the end of the cross-validation procedure, there is number of results equal to the number of cross-validation folds chosen for the procedure. Then, a best model must be chosen from these cross-validation results.

There are two ways of choosing this best model: one can choose Cross-Validation Consistency (CVC) or the lowest prediction error among the cross-validation models. CVC is the number of times the same model is chosen (based on lowest prediction error among all models) out of all rounds of cross-validation. The rationale for this is that if the model can indeed distinguish high-risk and low-risk patients, then it is likely to be found in a random subset of the data, similarly to how it would be expected to be found in separate, independent datasets. An issue arises when CVC is low, though (e.g., 1/5 or 2/5); multiple best cross-validation models may score the same, leading to ambiguity. Some MDR software will default to the first SNP which appears in the input dataset as a tiebreaker, potentially introducing systematic error bias into a study – as I show in my work described in Chapter 2. The other option is to choose the highest testing score, which represents the model most capable of distinguishing high-risk and low-risk patients in the independent testing set. This method has the advantage of having a much larger distribution of possible scores, and therefore it being unlikely that two or more scores will tie as the best choice. While there is still a potential for a result chosen by selecting the top scoring model to be spurious, regardless of the method used, this can be mitigated using a permutation testing procedure (see **Section 1.4.3**.).

1.4.2 Variations of MDR

As MDR was initially limited to certain study designs, as well as to univariate modelling, it has been expanded upon and modified by several groups to make it applicable for a variety of study designs and research questions. These methods vary in terms of implementation, methodology, and in the types of studies and data to which they are applicable. The work for this thesis focused on two MDR variations: GMDR (specifically the GMDR 0.9 implementation) and Cox-MDR.

1.4.2.1 GMDR

A notable extension of MDR is Generalized Multifactor Dimensionality Reduction (GMDR[101]). While the original MDR algorithm is limited to a small number of study designs (e.g., balanced case-control studies and discordant sib-pair studies), GMDR was designed to use generalized linear models (e.g., linear regression or logistic regression), greatly expanding the versatility of the software. As one can adjust for covariates in these models, multivariate GMDR analyses thus also became possible with this addition.

GMDR implements generalized linear models by utilizing a different scoring scheme for models than that of MDR. Instead of the case-control ratios of MDR, GMDR utilizes a score based on the generalized linear model (e.g., the logit score from a logistic regression model[101]). As in MDR, this score, integrated for all samples for each combination of possible genotypes for each combination of SNPs, is then used to determine whether this combination of SNP genotypes is high risk or low risk regarding

the response variable. GMDR also uses Balanced Accuracy (BA), in place of MDR's prediction error calculation. BA is calculated as the average of the sensitivity and the specificity of the model for its ability to predict response variable[121].

GMDR has previously been applied to cancer genetics to study SNP interactions. These studies, however, have predominantly studied susceptibility and risk, but have not been survival studies. For example, Wang et. al. applied GMDR to Wnt/β-Catenin genes to study cervical cancer susceptibility[122], Yu et al. applied GMDR to inflammation-related gene SNPs to study colorectal cancer risk[123], and Fu et al. applied GMDR to VEGF genes to study bladder cancer risk, finding several SNP-environment interactions[124]. Neither of these three papers identified any multi-loci models, however. On the other hand, Yadav et al. found a significant association between rs11954856 in the *APC* gene and rs4791171 in the *AXIN2* gene and increased susceptibility for gallbladder cancer[125]. Another study found several significant 4-5 loci models associated with colorectal cancer risk using GMDR, collectively involving SNPs from all three genes studied, *TACR1*, *TAC1*, and *TACR2*[126]. For treatment response, the interaction of polymorphisms CYP3A5*3, *NQO1* 609C>T, and *ABCB1* 1236C>T was found to be significantly associated with response to neoadjuvant chemotherapy in breast cancer patients, and interactions were also identified for association with anemia, leucopenia, and dosage delay/reduction due to development of fever[127]. These and other results published in the literature demonstrate the utility of GMDR in helping to resolve the missing heritability problem.

1.4.2.2 Cox-MDR

Cox-MDR is the 2nd MDR algorithm that I have used in my thesis research. While the GMDR algorithm was designed to work with any generalized linear model, practically, only specific distributions were implemented (e.g., linear regression, logistic regression). As such, other research groups developed methods derived from GMDR, utilizing different models, under different names. Cox-MDR is one such specification.

Cox-MDR employs the use of the Cox regression statistical method[105]. Martingale residuals from the Cox model are used as a means of scoring survival of samples and the sum of these residuals is used to determine whether a combination of variables is high-risk or low risk for poor survival. In survival studies, the utilization of the Cox regression method has a clear advantage over the use of linear regression or logistic regression: Cox regression is a method developed specifically for survival analysis, and thus Cox-MDR is able to utilize a full distribution of survival-time data and account for censored data. Therefore, Cox-MDR can be very valuable in examining SNP interactions in datasets where longitudinal survival times are utilized. Prior to my study, Cox-MDR had been successfully applied to other cancers to examine the prognostic associations of SNP interactions[105].

1.4.3 Permutation testing

MDR can identify interaction models. However, even with cross-validation being employed, there is still the possibility of spurious MDR results: the initial partitioning of the data into cross-validation folds may be responsible for a false-positive result and it is

possible that a result chosen based on a best prediction score was found by chance. As such, a permutation testing procedure is performed to ensure that the result found is not likely to have been found spuriously. For the Monte Carlo-type permutation testing procedure like that was used in my research, the data is shuffled such that the genetic data entries for each patient no longer correspond to the correct patient survival information. Effectively, this simulates a new dataset, based on the original data, for which the link between genetics and disease outcome is randomized. It is expected that there are no true correlations to be found in this simulated data and thus the highest scoring model in this dataset will be much worse at stratifying patients than a true model found with the correct associations intact. The permutation process is repeated multiple (e.g., 1,000) times and on each of these simulated datasets the MDR procedure is performed for the top model SNPs found in the original procedure. This permutation testing procedure results in a distribution of prediction scores, one for each permutation of the dataset.

A permutation testing p-value is determined by the proportion of prediction scores as great, or greater than, the p-value selected as the best model. If the original model represents an actual statistical effect, and not one found via multiple testing, it is expected to score highly in this distribution—the permutation p-value will be low. Considering the importance of permutation testing in reducing false-positive findings, I applied this approach to my results. The MDR models that are significant by permutation testing then can be examined by inference methods, such as Cox regression and logistic regression methods, to further assess how well the model can differentiate patients based on their outcome risk, as I have also done in my research.

In summary, despite the many challenges associated with examining and interpreting SNP interactions, identifying interactions can also be helpful in understanding and predicting the variable basis of clinical outcomes among patients. This was the main motivation to conduct this exciting research for my thesis.

1.5 **Rationale and objectives**

Colorectal cancer is a common disease with moderate survival rates. Identifying prognostic markers - including genetic markers - can increase the prognostic certainty, and, hence survival outcomes of patients. While there have been extensive studies, including GWASs, performed to examine the relationships of SNPs with survival times in colorectal cancer patients, these studies examine the relationship between single SNPs and survival outcomes individually. Hence, they miss the potential SNP interactions that may be associated with the outcome measures.

In this thesis research, I aimed to assess and apply MDR-based methods to examine associations of 1 to 3 way SNP interactions with survival outcomes in colorectal cancer. These SNPs were chosen from genes that are biologically relevant to disease progression: specifically, the MMP genes and genes of the *VEGF* pathway along with genes whose protein products are known to interact with them[28–31].

To do so, my specific objectives were to:

1) Compare the functionality and feasibility of two MDR-based methods, Cox-MDR and GMDR 0.9. This was achieved through the application of these methods to the data of a cohort of colorectal cancer patients, to examine the interactions

between MMP gene SNPs in relation to patient overall survival times

2) Apply these methods to seven datasets of SNPs from the VEGF ligand and receptor interaction networks to examine interactions between SNPs for disease specific survival times in a similar cohort of colorectal cancer patients to that of objective 1

I am excited to say that by a wide margin, my work is the largest SNP interaction study in colorectal cancer survival to date, exploring several orders of magnitude more interactions than had previously been studied. My results thus contribute to the scientific literature on biomarker identification, colorectal cancer, and SNP interactions significantly.

Chapter 2: Manuscript: Examining SNP-SNP interactions and risk of clinical outcomes in colorectal cancer using Multifactor Dimensionality Reduction based methods

A version of this chapter was published in Frontiers in Genetics, Cancer Genetics and Oncogenomics section (2022). https://doi.org/10.3389/fgene.2022.902217[128].

Supplementary information for this chapter is located in Appendix 1.

For a list of author contributions, place see the Co-authorship Statement on page xvii.

2.1 Authors and affiliations

Aaron Curtis[1,2], Yajun Yu[1,2], Megan Carey[1], Patrick Parfrey[3], Yildiz E. Yilmaz[1,3,4], Sevtap Savas[1,2,5*]

[1]Discipline of Genetics, Faculty of Medicine, Memorial University, St. John's, NL, Canada

[2]Division of Biomedical Sciences, Faculty of Medicine, Memorial University, St. John's, NL, Canada

[3]Discipline of Medicine, Faculty of Medicine, Memorial University, St. John's, NL, Canada

[4]Department of Mathematics and Statistics, Faculty of Science, Memorial University, St. John's, NL, Canada

[5]Discipline of Oncology, Faculty of Medicine, Memorial University, St. John's, NL, Canada

2.2 Abstract

Background: SNP interactions may explain the variable outcome risk among colorectal cancer patients. Examining SNP interactions is challenging, especially with large datasets. Multifactor Dimensionality Reduction (MDR)-based programs may address this problem.

Objectives: 1) To compare two MDR-based programs for their utility; and 2) to apply these programs to sets of MMP and VEGF-family gene SNPs in order to examine their interactions in relation to colorectal cancer survival outcomes.

Methods: This study applied two data reduction methods, Cox-MDR and GMDR 0.9, to study one to three way SNP interactions. Both programs were run using a 5-fold cross validation step and the top models were verified by permutation testing. Prognostic associations of the SNP interactions were verified using multivariable regression methods. Eight datasets, including SNPs from MMP family genes (n=201) and seven sets of VEGF-family interaction networks (n=1,517 SNPs) were examined.

Results: ~90 million potential interactions were examined. Analyses in the MMP and VEGF gene family datasets found several novel 1- to 3-way SNP interactions. These interactions were able to distinguish between the patients with different outcome risks (regression p-values 0.03–2.2E-09). The strongest association was detected for a 3-way interaction including *CHRM3*.rs665159_*EPN1*.rs6509955_*PTGER3*.rs1327460 variants.

Conclusion: Our work demonstrates the utility of data reduction methods while identifying potential prognostic markers in colorectal cancer.

2.3 Background

Colorectal cancer is a common disease accounting for ~10% of the global cancer cases[1]. The first years following diagnosis are critical and associated with a higher risk of negative disease outcomes[44]. Select disease, tumor, and patient characteristics[129–131] are helpful while estimating prognosis and making treatment recommendations. Sadly, the survival rates vary across different countries and a significant portion of the patients are lost to this disease (5-year survival rate ~<60%)[132–134]. In the current era of Personalized Medicine, one of the main aims is to identify additional prognostic markers that can help with better risk classification and improve patient outcomes.

Genetic variants, such as Single Nucleotide Polymorphisms (SNPs), are widely studied in prognostic research in oncology[70,83,135]. A common goal of this research area is to assess whether genetic variants are associated with, and hence, can be a marker of patient outcome risk. Survival studies examining genetic variants in colorectal cancer, including large-scale association studies[64,73,75,83–85] have mostly focused on analysis of SNPs one by one, assuming their individual effects and/or associations with the outcomes. This approach, while quite valuable, has also an obvious limitation: it misses detection of potential interactions among the variants.

It is possible that genetic variations jointly, but not alone, affect patient survival outcomes (i.e. interactions). That means that the effects of variants/genotypes are only detectable when they exist together in the patient genomes and are examined using

specific approaches. While it is possible to examine interactions using statistical methods, these analyses may suffer from several well-known complexities (e.g. sparse data, need for computational resources), especially as the number of variables examined increases[100]. As an example of this complexity, the number of possible combinations of three SNPs, or "3-way interactions", in a dataset of 100 SNPs is 161700, a large number of variables to study. Because of such methodological restrictions and the fact that there are large numbers of genetic variations in the human genome, it is necessary to apply other approaches, such as data reduction methods, for comprehensive SNP interaction analyses. Multifactor Dimensionality Reduction (MDR) is a data reduction method designed for use in studies examining the interactions among variables while accounting for difficulties inherent in interaction analysis[99]. Initially created to support a small number of study designs, MDR has since been adapted for other types of studies. Generalized MDR (GMDR)[101] is an extension of MDR to support generalized linear models (e.g. logistic regression). Cox-MDR[105] is a type of GMDR which is designed specifically for survival/time-to-event studies and utilizes the Cox-regression method.

Studies that have so far considered the interactions of genetic variants in colorectal cancer outcomes using MDR are quite limited[94–98,136]. As a result, potential SNP interactions that may be associated with patient outcomes largely remain unknown. In this study, we aimed to explore the potential roles of SNP interactions in outcome risk of colorectal cancer patients using MDR-based methods. For this purpose, we utilized the genotype and outcome data of a cohort of colorectal cancer patients from Newfoundland and Labrador. We explored and compared the functionality of two MDR-based software

– Cox-MDR[105] and GMDR 0.9[101], and applied these software to examine the interactions among SNPs from the Matrix Metalloproteinase (MMP*)* family of genes and Vascular Endothelial Growth Factor (VEGF*)*-family interaction network genes. Our results show that there are unique limitations and strengths of Cox-MDR and GMDR 0.9, which should be considered in future studies. More importantly, our results identified novel SNP interactions that can help distinguish between colorectal cancer patients with significantly different outcome risks.

2.4 **Data and Methods**

2.4.1 Ethics approval

This study was conducted with ethics approval by the Health Research Ethics Authority of Newfoundland and Labrador (HREB #2018.051; #2009.106). This study was a secondary use of data study, hence, HREB waived the requirement for patient consent.

Part 1: ***Exploration of Cox-MDR and GMDR 0.9 programs and analysis of interactions between the SNPs from the MMP family of genes***

2.4.2 Patient cohort, genes selected, outcome measures, covariates, and data considerations

This is a cohort study. The baseline characteristics of the patient cohort included in this part of the study (n=439) are shown in **Appendix 1-Table S1**. Patients were

recruited by the Newfoundland Familial Colorectal Cancer Registry (NFCCR)[137,138]. They were under the age of 76 at the time of diagnosis and were diagnosed with colorectal cancer between 1999 and 2003. Pathological/clinical and follow-up data were collected from resources such as clinical reports, the Newfoundland Cancer Treatment and Research Foundation database, and follow-up questionnaires[44,71,137,138]. The date of last follow up was 2010. Genetic data was previously obtained from blood samples via the Illumina Omni1-Quad human SNP genotyping platform (reactions were outsourced to Centrillion Biosciences, United States), and sample quality control (QC) measures were implemented[83]. As a result, all patients included into the analyses were of Caucasian ancestry and unrelated to each other (i.e., not first, second, or third degree relatives[83]).

Since one of our aims in Part 1 was to examine and compare the performance and functionality of the two MDR-based programs, we opted for a set of genes and SNPs that were previously examined in our lab (**Appendix 1 – Table S2**). Specifically, the best suited genetic model for SNPs from the MMP genes and their one-by-one associations with patient outcomes were previously examined[33]. This previous knowledge enabled us to assess the results of the 1-way interaction analyses obtained using the MDR methods during the current study, by comparing them to the results obtained in the previous study. We kept the covariates and outcome measure examined in Part 1 the same as in that previous study. The covariates included age at diagnosis, disease stage, MSI (microsatellite instability)-status, and tumor location (rectum, colon). The outcome of interest was death from any cause (Overall Survival; OS).

Since Cox-MDR and GMDR 0.9 make their calculations, classify the patient genotypes as high-risk or low-risk, and select best models based on different scoring methods (i.e. martingale residuals obtained by Cox regression in Cox-MDR and logit score obtained by logistic regression in GMDR 0.9), Cox-MDR and GMDR 0.9 differ in data requirements. For example, as GMDR 0.9 utilizes logistic regression method, the 5-year-survival outcome measure was used. In Cox-MDR analysis, survival status and time to death (or the last date of alive contact) were used. Considering these and additional input data requirements for each program, a number of measures were taken while preparing the data files for analysis (see **Appendix 1** for details). Since we aimed to compare their performance in this first part of the study, we also examined the same set of patients in the Cox-MDR and GMDR 0.9 analyses.

2.4.3 Single Nucleotide Polymorphism genotype data and quality control measures

SNPs from the MMP family genes were extracted from the genome-wide SNP genotype data files using the gene genomic location information and the PLINK software[139,140] (version 1.07), with the following quality control parameters being implemented: minor allele frequency (MAF) >= 0.05, Hardy-Weinberg Equilibrium (HWE) $p > 0.0001$, and missing genotype rate = 0. Pairwise squared correlation coefficient (r^2) values and MAFs were calculated using PLINK. When there were multiple SNPs with $r^2 = 1$ (i.e. those which would score identically using the MDR procedure), SNPs were removed such that only one of these SNPs was present in the final

dataset. As a result, 201 SNPs from 21 MMP genes were included into the analysis (**Appendix 1 - Table S2**).

2.4.4 Cox-MDR and GMDR 0.9 analyses

The work-flow is summarized in **Figure 2.1**.

We focused on 1-way, 2-way, and 3-way (k=1-3) interactions. 1-way interaction analysis examines whether the genotype groups of a single SNP may be categorized as high-risk and low-risk genotypes, and associated with an outcome/response variable. 2-way and 3-way interaction analyses examine whether combinations of genotype groups of two or three SNPs may be categorized as high-risk and low-risk genotypes, and associated with an outcome/response variable, respectively. Cox-MDR uses martingale residuals of Cox-regression models[105] and GMDR 0.9[101] uses logit scores to categorize patient genotypes as high-risk and low-risk genotypes.

Cox-MDR code[105] was requested and received from the developer, Dr. Seungyeoun Lee (Sejong University, South Korea). We extended the code in order to add additional functionality and return the output that would be needed for our study using R[141] (**Appendix 1**). GMDR 0.9 code was downloaded from the UAB Department of Biostatistics Section on Statistical Genetics website (GMDR) on December 11, 2018. Command line arguments to set the random seeds were added to the permutation testing Perl script included with GMDR 0.9 (**Appendix 1**). Once we verified that Cox-MDR

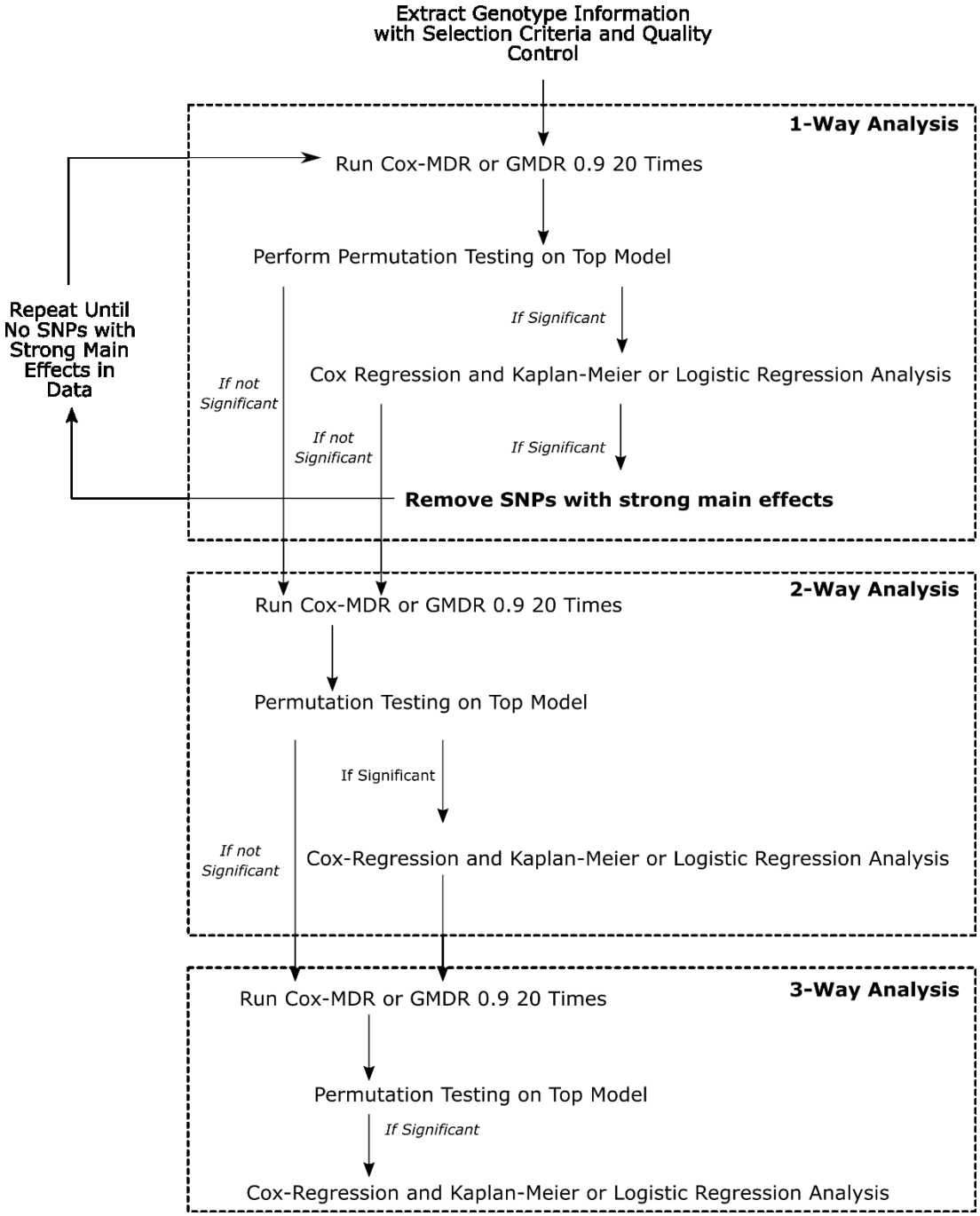


Figure 2.1: Overall workflow diagram for MDR analysis protocol. This figure demonstrates the overall workflow of the analyses performed. Multivariate Cox-regression and univariate Kaplan-Meier analyses were used to verify the Cox-MDR results and assess the associations of the identified genotype groups with clinical outcomes, whereas multivariate logistic regression was used to verify the GMDR 0.9 results and the association of the identified genotype groups with clinical outcomes.

worked as expected, it was run with the dataset (including both the clinical [i.e. covariates and OS time and status] and the genotype data of the SNPs from the MMP genes).

All interaction analyses were performed using a 5-fold cross-validation procedure. 5-fold cross validation is appropriate when the sample size is modest, like ours, while still providing adequate power[142]. 4/5 of these folds served as a training set for the MDR procedure and the final 1/5 was an independent testing set from which the final model score was derived. The code was run 20 times, each run yielding a "best Cox-MDR model", with different random seeds to ensure different partitioning of the dataset into each of the 5 cross-validation folds (i.e. to reduce the influence of any specific partitioning of the data). Given the 5-fold cross-validation procedure, this resulted in each SNP or SNP combination being examined in potentially a total of 100 patient datasets. Among the best Cox-MDR models returned by each of the 20 runs, we prioritized the most frequently detected best Cox-MDR model (with consistent SNP ID(s) and high-risk and low risk genotype information) with the highest testing balance accuracy (TBA) score. We refer to these models as the "top" Cox-MDR models throughout this manuscript.

GMDR 0.9 was applied to the same dataset as used in Cox-MDR, with the only exception of using the 5-year survival status as the response variable. In contrast to Cox-MDR, GMDR 0.9 can only select the best models based on the cross-validation consistency (CVC); that is, the model with the highest CVC among cross-validation folds is selected. After running the GMDR 0.9 analysis 20 times, we selected the top model as

in Cox-MDR and based on the highest average TBA value among cross validation folds (GMDR 0.9's analogue to Cox-MDR's highest TBA). In cases when there were multiple models satisfying the best MDR model criteria in a dataset, we used the TBA, and if still needed, the CVC information, as the tie breaker.

2.4.5 Permutation testing

Once the top Cox-MDR or GMDR 0.9 model was identified, the significance of the model was assessed using permutation testing. For GMDR 0.9, permutation testing was performed using the included Perl script, which was extended to allow setting of random seeds. For Cox-MDR permutation testing, an R function was written. The permutation procedure was performed using 1,000 permutations of the data (**Appendix 1**).

Permutation testing was performed for all top models selected from k-way runs (1-3-ways). As noted by others[99,105,142–146], it is possible that a single SNP with a strong main effect (that can be identified as the top MDR-model in 1-way analysis), may impact higher order interaction analysis when using MDR-based methods, and hence, needs to be removed from the 2-way and 3-way interaction analyses. Therefore, we first performed the permutation testing for the top MDR model identified in the 1-way analysis and, if it turned out to be a significant MDR model, then we assessed whether the high-risk and low-risk genotype groups of this top model were associated with survival outcomes in the patient cohort using statistical methods (see below). In the case where a significant association was detected, we then performed subsequent runs by

excluding this SNP and any other SNP in the dataset that was in high linkage disequilibrium (LD) with it ($r^2 \geq 0.8$). This SNP removal procedure was repeated until all SNPs with strong main effects in 1-way analyses were removed from the dataset (**Figure 2.1**). We then proceeded to 2-way and 3-way analyses on the final dataset with all SNPs with strong main effects removed.

2.4.6 Kaplan-Meier curves and multivariable regression analyses

Following identification of a significant top MDR model by permutation testing, we assessed whether the high-risk and low-risk genotype groups of the model were associated with survival outcomes in the patient cohort. For this purpose, we applied multivariable Cox regression analysis (for the models identified by Cox-MDR) and logistic regression analysis (for the models identified by GMDR 0.9) using the same clinical covariates for adjustment that were used in the Cox-MDR and GMDR 0.9 runs. When needed, Kaplan-Meier curves were constructed to visualize the survival times of the patient groups with the high-risk and low-risk genotype groups over time. These analyses were performed using IBM SPSS Statistics software (versions 25 and 26, Armong, NY)[147] or R. A p-value of < 0.05 was considered significant.

Part 2: ***Interactions among the SNPs of the VEGF interaction networks***

Data resources and methods for Part 2 of this study were similar to Part 1, except for the differences outlined in this section. Four hundred patients (**Appendix 1 - Table S3**) met the data requirements. All 400 of these patients were used in the Cox-MDR

analysis. For Cox-MDR analysis, Disease Specific Survival (DSS) was used as an outcome measure, where the endpoint was death from colorectal cancer. For GMDR 0.9 analysis 5-year DSS time was used as the outcome measure. Using this outcome measure, five patients, who were censored prior to 5 years were excluded from analysis, as the survival status of these patients at 5 years was unknown. This left 395 patients for analysis with the GMDR 0.9 algorithm. An updated outcome data (with the last follow-up date of 2018[44]) was used in this part of the study. Clinical variables that were previously identified as prognostic markers for DSS[44] were used as covariates in Cox-MDR, GMDR 0.9, and Cox regression and logistic regression analyses (tumor location, stage, MSI status, adjuvant chemotherapy and radiotherapy status).

For this part of the study, we focused on the VEGF family members and examined SNP interactions in their protein-protein interaction networks. Four ligands (*VEGFA*, *VEGFB*, *VEGFC*, and *PIGF*) and three receptors (*VEGFR1*, *VEGFR2*, and *VEGFR3*) were selected. Since association studies using the sex chromosome genetic variations face additional complexities, the fifth ligand, *VEGFD*, which is located on the X chromosome, was not included.

2.4.7 Identification of interaction partners of the VEGF family proteins

Each of the seven VEGF proteins were searched in the BioGRID 3.5 database[148–150] to find proteins that interact with them (i.e. protein-protein interaction networks; BioGRID accessed on October 22, 2019). Genomic locations for all interactors were obtained from the Ensembl database[151,152](GRCh37 assembly) using the legacy archive

Biomart[153]. PLINK was used for genotype extraction from the genome-wide SNP genotype data files, followed by LD-based pruning. Interactors located on the X chromosome (*FIGF*, *IKBKG*, and *VSIG4*) and genes with no SNPs after quality control and pruning steps (*BCS1L*, *CTGF*, *LRFN3*, *NUDT16L1*, *SCH1*, *TXNIP*, and *UBIAD1*) were excluded. In 7 VEGF networks, there was a total of 1,517 unique SNPs (number of SNPs in each set: *VEGFA*=401; *VEGFB*=174; *VEGFC*=38; *PIGF*=102; *VEGFR1*=222; *VEGFR2*=747; *VEGFR3*=328) in a total of 131 unique genes (number of genes in each set: *VEGFA*=43; *VEGFB*=14; *VEGFC*=3; *PIGF*=5; *VEGFR1*=15; *VEGFR2*=68; *VEGFR3*=23). Please see **Appendix 1 Figure S1** and **Appendix 1 Tables S4-S5** for the interaction networks, proteins in each interactome, and the IDs of SNPs retrieved and analyzed in this part of the study.

2.4.8 Bioinformatics analyses

In order to explore the links between the SNPs of interest and clinical outcomes, we utilized literature reports (from PUBMED), and dbANGIO[154] and dbCPCO[69] databases. We also searched RegulomeDB[155,156] (accessed on September 20, 2019, November 5, 2020) and GTEx databases[157] (accessed on November 15, 2020) to identify eQTLs that are associated with expression levels of genes (Note that GTEx has no data for rectal tissues, so only transverse and sigmoid colon tissue information was available). Information on the type of variation (e.g. intronic) were retrieved from dbSNP[158].

2.5 **Results**

Part 1: Examination of the interactions between the MMP gene family SNPs using Cox-MDR and GMDR 0.9

Interactions among 201 SNPs from 21 MMP genes were examined as a set (a total of 1,353,601 potential interactions). As a result, 1-way Cox-MDR interaction analysis identified *MMP27*-rs11225388 (MAF = 0.27; an intronic SNP) and classified its genotypes as high-risk (AA) and low-risk (AG and GG) in the top MDR model. Permutation testing was also significant (p = 0.011). It is interesting that the best MDR-models identified by each of the 20 individual runs identified this SNP and its genotype categories consistently (**Appendix 1 – Table S6**). Multivariable Cox regression analysis, adjusting the rs11225388 genotypes (low risk genotypes versus high risk) for clinical covariates, showed that this SNP genotype model was independently associated with OS (**Table 2.1**). Therefore, Cox-MDR successfully identified a significant 1-way interaction. These results also meant that the rs11225388 SNP had a significant main effect, which necessitated it (as well as two other SNPs with high LD with it: rs11225389 and rs12365082) being removed from the dataset prior to future analyses. Upon re-running Cox-MDR 1-way analysis and applying permutation testing to the top model, we did not identify a significant 1-way MDR model. We, therefore, proceeded with 2-way and 3-way analysis. These runs did not identify any significant multi-loci Cox-MDR models in this dataset.

Table 2.1: Multivariable Cox regression analysis result for the significant 1-way Cox-MDR model in the MMP dataset (overall survival).

Top Model SNP	**High risk genotypes**	**p-value**	**HR**	**95% CI (lower-upper)**
rs11225388_GA	AA	0.002	0.591	0.425-0.821

CI: confidence interval; HR: hazards ratio; SNP: single nucleotide polymorphism. HR calculated for low risk genotypes (GG+GA) versus high-risk genotype (AA).

In contrast, in the 1-way analysis, GMDR 0.9 selection procedure did not identify a significant model following permutation testing. However, 2-way analysis identified a two-loci MDR model including the *MMP16*.rs7817382 and *MMP24*.rs2254207 variants (permutation testing p=0.001; **Table 2.2**). Multivariable logistic regression analysis verified that this model had a significant association with 5-year survival of patients when adjusted for other prognostic covariates (high risk genotypes versus low risk genotypes; OR: 3.27; p=4E-6). Both of these SNPs are non-coding region SNPs and were common in the patient cohort (MAFs = 0.25 and 0.26, respectively). Additionally, in the 3-way analysis, a GMDR 0.9 model including genotypes of *MMP16*.rs2664369, *MMP20*.rs11225332, and *MMP2*.rs11639960 variants were identified in the top model (permutation testing $p < 0.001$). Multivariable logistic regression analysis showed that this model distinguished patients based on their 5-year survival status independent of other covariates and this association was quite strong (p=1.3E-8; OR: 4.5; **Table 2.2**). Kaplan Meier curves for the identified high-risk and low-risk genotypes are shown in **Appendix 1 – Figure S2**. Rs2664369 is a 3'-untranslated region variant, and rs11225332 and rs11639960 are both intronic variants. These SNPs were common in the patient cohort (MAF = 0.43, 0.40, and 0.35, respectively).

Table 2.2: Multivariable logistic regression analysis results for the significant 2-way and 3-way GMDR 0.9 models in the MMP dataset (overall survival).

Top Model SNPs	**High risk genotypes**	**p-value**	**OR**	**95% CI (lower-upper)**
rs7817382_GA and rs2254207_CA	(0AA,1CA),(0AA,2CC),(1GA,0AA),(1GA,2CC),(2GG,1CA)	4.4194E-06	3.266	1.971-5.414
rs2664369_GT, rs11225332_CT and rs11639960_GA	(0TT,0TT,2GG),(0TT,1CT,1GA),(0TT,1CT,2GG),(0TT,2CC,1GA),(1GT,0TT,0AA),(1GT,0TT,1GA),(1GT,1CT,2GG),(1GT,2CC,2GG),(2GG,0TT,0AA),(2GG,1CT,2GG),(2GG,2CC,0AA),(2GG,2CC,2GG)	1.2929E-08	4.503	2.681-7.563

CI: confidence interval; OR: odds ratio; SNP: single nucleotide polymorphism.
Alleles are given in the order major allele minor allele. 0,1,2, refer to additive coding, i.e. dosage of the minor allele. (0 = 0 copies of the minor allele, 1 = 1 copies of the minor allele, 2 = 2 copies of the minor allele).

Part 2: Examination of the interactions in the VEGF interaction network datasets using Cox-MDR and GMDR 0.9

In this part of the study, we investigated SNP interactions separately for seven sets of VEGF family protein interaction networks (**Appendix 1 - Tables S4-S5**). Altogether, these analyses examined 88,989,448 potential interactions.

Cox-MDR identified four significant MDR models, three of which were also confirmed by multivariable Cox regression analysis (**Table 2.3**). In the 1-way analysis of the *PIGF* network, we identified one SNP associated with DSS (*RNF123*.rs11130216). Additionally, both 2-way and 3-way interactions were detected and they were both identified during the *VEGFR3* network analysis. These multi-loci interactions include SNPs from *CHRM3*, *PTGER3*, or *EPN1* genes. The strongest association with disease-specific survival was detected in the 3-way analysis with a very strong p-value of 2.21E-09 (*CHRM3*.rs665159_*EPN1*.rs6509955_*PTGER3*.rs1327460; HR: 5.0). As also demonstrated by the Kaplan Meier curve (**Fig. 2.2**), this model's genotype classification was able to clearly separate patients based on their outcome risks.

Table 2.3: Permutation testing and multivariable Cox-regression analysis results for the top Cox-MDR models in the VEGF interaction network set analyses (disease specific survival).

Interactor Set	**Top model SNP(s)**	**High risk genotypes**	**Permutation p-value**	**Cox regression p-value**	**HR**	**95% CI (lower-upper)**
1-way						
Iteration 1						
VEGFA	*FN1*.rs2289200[TG]	1(TG),2(TT)	0.273	--	--	--
VEGFB	*VEGFA*.rs833070[GA]	1(GA)	0.201	--	--	--
VEGFC	*VEGFC*.rs1485766[CA]	1(CA)	0.346	--	--	--
VEGFR1	*PIK3R1*.rs4122269[CT]	0(TT)	0.07	--	--	--
VEGFR2	*PTPN12*.rs1024723[TC]	0(CC),2(TT)	0.181	--	--	--
VEGFR3	*LRRK1*.rs930847[CA]	1(CA),2(CC)	0.098	--	--	--
PIGF	*RNF123*.rs11130216[AC]	1(AC),2(AA)	0.032	0.003	1.977	1.265-3.089
Iteration 2						
PIGF	*VEGFA*.rs833070[GA]	1(GA)	0.045	0.298	1.256	0.818-1.928
2-way						
VEGFA	*CLU*.rs7982[TC], *FLT1*.rs7332329[GA]	(0[CC],0[AA]),(1[TC],1[GA])(0[CC],2[GG])(2[TT],2[GG])	0.392	--	--	--
VEGFB	*FAT1*.rs10155467[TC], *VEGFA*.rs3025010[CT]	(1[TC],0[TT])(0[CC],1[CT])(2[TT],1[CT])(0[CC],2[CC])(2[TT],2[CC])	0.225	--	--	--

VEGFC	*KDR*.rs17709898[GA], *VEGFC*.rs3775195[AC]	(0[AA],0[CC])(2[GG],0[CC])(1[GA],1[AC])(0[AA],2[AA])(1[GA],2[AA])	0.146	--	--	--
VEGFR1	*FLT1*.rs9551462[TC], *PIK3R1*.rs1823023[AG]	(1[TC],0[GG])(2[TT],0[GG])(0[CC],1[AG])(0[CC],2[AA])(1[TC],2[AA])	0.128	--	--	--
VEGFR2	*APP*.rs2096488[CA], *DNM2*.rs7246673[TG]	(2[CC],0[GG])(0[AA],1[TG])(1[CA],2[TT])(2[CC],2[TT])	0.389	--	--	--
VEGFR3	*CHRM3*.rs665159[TC], *PTGER3*.rs1327460[AG]	(0[CC],0[GG])(1[TC],0[GG])(0[CC],1[AG])(1[TC],2[AA])	0.004	2.03E-06	3.147	1.961-5.050
PIGF	*NRP1*.rs2474723[GA], *RNF123*.kgp9864706[AG]	(0[AA],0[GG])(1[GA],2[AA])	0.527	--	--	--
3-way						
VEGFA	*FOS*.rs7101[CT], *NRP2*.rs861079[TC], *TFAP2A*.rs303055[CT]	(0[TT],0[CC],0[TT])(1[CT],0[CC],0[TT])(0[TT],1[TC],0[TT])(0[TT],2[TT],0[TT])(0[TT],0[CC],1[CT])(1[CT],1[TC],1[CT])(2[CC],1[TC],1[CT])(0[TT],2[TT],1[CT])(1[CT],2[TT],1[CT])(2[CC],2[TT],1[CT])(0[TT],0[CC],2[CC])(2[CC],0[CC],2[CC])(1[CT],2[TT],2[CC])	0.058	--	--	--
VEGFB	*ALOXE3*.rs3809882[CA], *COL6A2*.rs7280485[AG], *NRP1*.rs6481844[CT]	(1[CA],0[GG],0[TT])(0[AA],1[AG],0[TT])(2[CC],1[AG],0[TT])(1[CA],2[AA],0[TT])(0[AA],0[GG],1[CT])(2[CC],0[GG],1[CT])(0[AA],1[AG],1[CT])(1[CA],2[AA],1[CT])(2[CC],2[AA],1[CT])(1[CA],1[AG],2[CC])(2[CC],2[AA],2[CC])	0.217	--	--	--

VEGFC	*FLT4*.rs2242217[CT], *FLT4*.rs11748431[AG], *VEGFC*.rs1485762[TC]	(2[CC],0[GG],0[CC])(0[TT],1[AG],0[CC])(2[CC],1[AG],0[CC])(1[CT],2[AA],0[CC])(1[CT],0[GG],1[TC])(2[CC],0[GG],1[TC])(1[CT],1[AG],1[TC])(2[CC],1[AG],1[TC])(1[CT],2[AA],1[TC])(0[TT],0[GG],2[TT])(2[CC],0[GG],2[TT])(1[CT],1[AG],2[TT])	0.229	--	--	--
VEGFR1	*FLT1*.rs12429309[CT], *FLT1*.rs9551462[TC], *PIK3R1*.rs1823023[AG]	(1[CT],0[CC],0[GG])(1[CT],1[TC],0[GG])(2[CC],1[TC],0[GG])(0[TT],2[TT],0[GG])(0[TT],0[CC],1[AG])(0[TT],0[CC],2[AA])(2[CC],0[CC],2[AA])(0[TT],1[TC],2[AA])(1[CT],1[TC],2[AA])	0.097	--	--	--
VEGFR2	*COL18A1*.rs4819101[AG], *NCOA4*.rs10761581[GT], *PALLD*.rs10004025[TC]	(0[GG],0[TT],0[CC])(1[AG],0[TT],0[CC])(2[AA],0[TT],0[CC])(0[GG],1[GT],0[CC])(2[AA],0[TT],1[TC])(0[GG],1[GT],1[TC])(0[GG],2[GG],1[TC])(1[AG],2[GG],1[TC])(1[AG],0[TT],2[TT])(2[AA],0[TT],2[TT])(1[AG],2[GG],2[TT])(2[AA],2[GG],2[TT])	0.12	--	--	--
VEGFR3	*CHRM3*.rs665159[TC], *EPN1*.rs6509955[AG], *PTGER3*.rs1327460[AG]	(1[TC],0[GG],0[GG])(0[CC],1[AG],0[GG])(1[TC],1[AG],0[GG])(0[CC],2[AA],0[GG])(0[CC],0[GG],1[AG])(0[CC],1[AG],1[AG])(0[CC],2[AA],1[AG])(1[TC],2[AA],1[AG])(1[TC],0[GG],2[AA])(2[TT],0[GG],2[AA])(1[TC],1[AG],2[AA])(0[CC],2[AA],2[AA])(2[TT],2[AA],2[AA])	0.007	2.21E-09	5.004	2.952-8.481

PIGF	*FLT1*.rs17086609[GA], *FLT1*.rs1853581[CA], *NRP1*.rs2506141[CT]	(1[GA],0[AA],0[TT])(0[AA],1[CA],0[TT])(0[AA],2[CC],0[TT])(2[GG],2[CC],0[TT])(1[GA],0[AA],1[CT])(0[AA],1[CA],1[CT])(2[GG],1[CA],1[CT])(0[AA],0[AA],2[CC])(2[GG],0[AA],2[CC])(2[GG],1[CA],2[CC])	0.253	--	--	--

CI: confidence interval; HR: hazards ratio; SNP: single nucleotide polymorphism.
0, 1, and 2 in the High Risk Genotype column refer to additive coding, where the number refers to the number of minor alleles in the genotype.
Square brackets in the Top Model SNPs column indicate major and minor alleles for each SNP; which the first letter represents the minor allele and the second letter represents the major allele. In the high risk genotypes column, the three items enclosed in parentheses signify the genotypes of the combination of SNPs which was found to be high risk by Cox-MDR. Commas separate the genotypes for each SNP in the order in which they appear in the corresponding Top Model SNPs entry. Whenever a SNP with a main effect was identified in 1-way analysis, the analysis was repeated with that SNP removed from the dataset (i.e. successive iterations). *FLT1* is also known as *VEGFR1*; *KDR* is also known as *VEGFR2*; *FLT4* is also known as *VEGFR3*; and *PGF* is also known as *PIGF*.

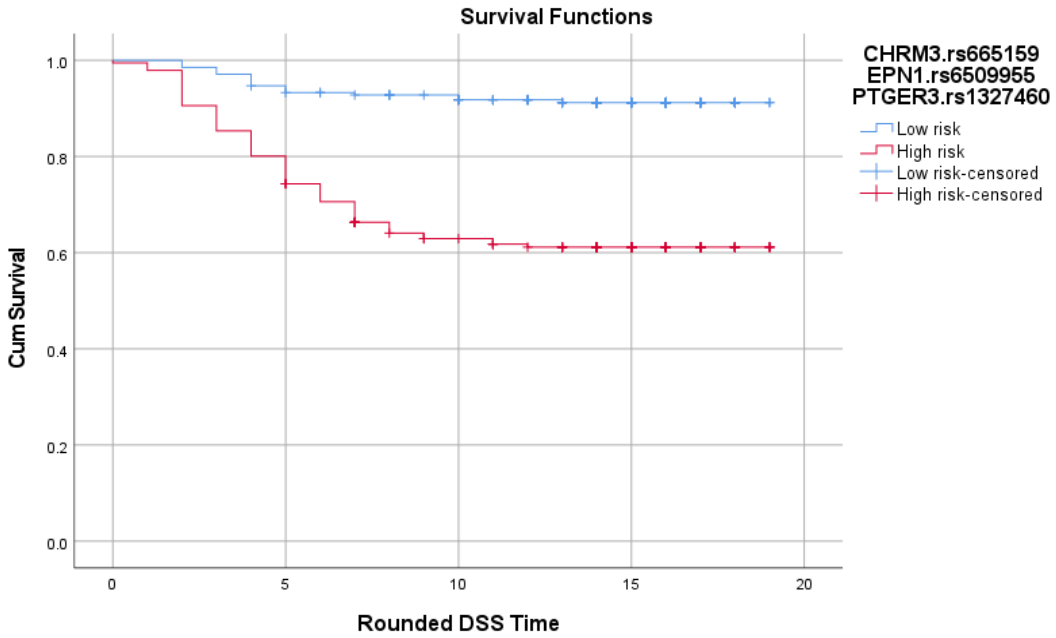


Figure 2.2: Kaplan-Meier curve for 3-way Cox-MDR analysis, VEGFR3 dataset. *Log-rank p = 1.02619688760668E-12. Red: high risk genotype combinations: (TC,GG,GG), (CC,AG,GG), (TC,AG,GG), (CC,AA,GG), (CC,GG,AG), (CC,AG,AG), (CC,AA,AG), (TC,AA,AG), (TC,GG,AA), (TT,GG,AA), (TC,AG,AA), (CC,AA,AA), (TT,AA,AA). Blue: all other genotype combinations. The vertical lines on the curves denote the censored patients (e.g. patients alive at the last follow up time). X and Y axis show the follow-up time (in years; rounded) and cumulative survival, respectively.*

Similar to Cox-MDR, GMDR 0.9 also identified interactions that were able to distinguish between patients with different outcome risk (the multivariable logistic regression p-values 0.032 – 2.4E-09; **Table 2.4**). GMDR 0.9 identified a larger number significant interactions than Cox-MDR (11, six, and seven 1-way, 2-way, and 3-way interactions, respectively). The strongest association with DSS (p=2.4E-09) was detected for the 3-way *ADRB2*.rs1042711_*NRP1*.rs17296436_*VEGFB*.rs11603042 interaction in the *VEGFB* network analysis (HR: 10, 95% CI: 4.691-21.276; Kaplan Meier curves for the high-risk and low-risk genotypes are shown in **Figure 2.3**). Overall, the significant associations, particularly for multi-loci interactions, were quite encouraging. Generally, the significance levels of interactions increased with the order of interactions (i.e. from 1-way to 3-way). Of note, 3-way analysis identified significant interactions in all seven VEGF interaction networks examined. Rarely, interaction models included both the VEGF ligand and receptor (*FLT4*.rs307823_*KDR*.rs6828477_*KDR*.rs12502008) or two SNPs from the same gene (*FLT4*.rs11739750_*FLT4*.rs307814; **Table 2.4**), both detected in the *VEGFC* interaction network. For interested readers, the Kaplan Meier curves for the GMDR 0.9 identified interactions are shown in **Appendix 1 – Figure S3**.

Table 2.4: Multivariable logistic regression analysis results for the top GMDR 0.9 models in the VEGF interaction network set analyses (disease-specific survival).

1-way						
Iteration 1						
Interaction Set	**Top model SNP(s)**	**High risk genotypes**	**Permutation p-value**	**Logistic regression p-value**	**OR**	**95% CI (lower-upper)**
VEGFA	*NRP2*.rs3771003[TG]	0[GG],2[TT]	0.014	0.010	2.399	1.230-4.679
VEGFB	*COL6A2*.rs9978018[GA]	0[AA],2[GG]	0.02	0.032	2.015	1.062-3.822
VEGFC	*FLT4*.rs3797102[CT]	1[CT],2[CC]	0.358	--	--	
VEGFR1	*MICAL2*.rs11022250[GT]	0[TT]	<0.001	0.002	2.941	1.468-5.891
VEGFR2	*PTPN12*.rs1024723[TC]	0[CC],2[TT]	<0.001	1.442E-04	3.662	1.875-7.152
VEGFR3	*CHRM3*.rs12037424[CT]	0[TT]	0.005	0.004	2.616	1.369-4.997
PIGF	*RNF123*.rs11130216[AC]	1[AC],2[AA]	0.045	0.011	2.359	1.222-4.554
Iteration 2						
VEGFA	*HNRNPL*.rs10403012[GA]	0[AA]	0.022	0.012	1.984	0.673-5.847
VEGFB	*VEGFB*.rs11603042[TG]	1[TG],2[TT]	0.067	--	--	
VEGFR1	*MICAL2*.rs988189[TC]	1[TC],2[TT]	0.116	--	--	
VEGFR2	*MAPK1*.rs2298432[AC]	0[CC]	0.001	3.425E-04	3.467	1.756-6.848
VEGFR3	*CHRM3*.rs2278642[TG]	1[TG],2[TT]	0.007	0.006	2.924	1.362-6.278
PIGF	*FLT1*.rs3936415[AG]	0[GG]	0.069	--	--	
Iteration 3						
VEGFA	*HNRNPL*.rs2278012[CT]	0[TT]	0.051	--	--	
VEGFR2	*DNM2*.rs7246673[TG]	1[TG],2[TT]	0.079	--	--	

VEGFR3	*LRRK1*.rs12595297[GT]	1[GT]	0.007	0.011	2.243	1.207-4.169
Iteration 4						
VEGFR3	*LRRK1*.rs17161155[AG]	0[GG]	0.043	0.009	2.317	1.235-4.346
Iteration 5						
VEGFR3	*CHRM3*.rs6692711[TC]	1[TC]	0.225	--	--	
2-way						
VEGFA *FLT1*.rs3936415[AG]	*ELAVL1*.rs3786619[AG]	(0[GG],2[AA])(1[AG],0[GG])(2[AA],0[GG])(2[AA],1[AG])	<0.001	3.180E-05	4.387	2.186-8.805
VEGFB *HAL*.rs3213737[CT]	*ADRB2*.rs1042711[CT]	(0[TT],1[CT])(1[CT],0[TT])(2[CC],1[CT])	0.018	7.082E-05	3.696	1.940-7.044
VEGFC *FLT4*.rs307814[TC]	*FLT4*.rs11739750[TC]	(0[CC],1[TC])(1[TC],0[CC])(1[TC],2[TT])(2[TT],1[TC])	0.002	1.335E-04	3.827	1.922-7.620
VEGFR1 *MICAL2*.rs7946327[CA]	*FLT1*.rs3794397[TC]	(0[CC],0[AA])(1[TC],1[CA])(2[TT],1[CA])	0.003	1.852E-04	3.361	1.780-6.346
VEGFR2 *PTPRR*.rs4760847[GA]	*COL18A1*.rs7278425[TC]	(0[CC],1[GA])(1[TC],0[AA])	<0.001	1.213E-05	4.542	2.306-8.947
VEGFR3 *TMEM52B*.rs10505752[TC]	*CHRM3*.rs1782357[TC]	(0[CC],0[CC])(1[TC],1[TC])(1[TC],2[TT])(2[TT],2[TT])	<0.001	3.872E-05	3.892	2.037-7.433
PIGF *NRP1*.rs12762312[TC]	*FLT1*.rs2387632[TC]	(0[CC],1[TC])(1[TC],0[CC])(1[TC],2[TT])(2[TT],0[CC])	0.055	--	--	
3-way						
VEGFA *ELAVL1*.rs3786619[AG] *NRP2*.rs861079[TC]	*CLU*.rs9331888[CG]	(0[GG],0[GG],1[TC])(0[GG],0[GG],2[TT])(0[GG],1[AG],1[TC])(0[GG],2[AA],0[CC])(0[GG	0.001	2.146E-07	9.322	4.010-21.672

		],2[AA],2[TT])(1[CG],1[AG],0[CC])(1[CG],1[AG],2[TT])(1[CG],2[AA],0[CC])(2[CC],0[GG],1[TC])(2[CC],0[GG],2[TT])(2[CC],1[AG],2[TT])(2[CC],2[AA],2[TT])				
VEGFB	*ADRB2*.rs1042711[CT] *NRP1*.rs17296436[GA] *VEGFB*.rs11603042[TG]	(0[TT],0[AA],1[TG])(0[TT],0[AA],2[TT])(0[TT],1[GA],2[TT])(0[TT],2[GG],1[TG])(1[CT],0[AA],2[TT])(1[CT],1[GA],0[GG])(1[CT],2[GG],0[GG])(1[CT],2[GG],1[TG])(2[CC],1[GA],0[GG])(2[CC],1[GA],1[TG])(2[CC],1[GA],2[TT])	0.007	2.404E-09	9.991	4.691-21.276
VEGFC	*FLT4*.rs307823[GA] *KDR*.rs6828477[CT] *KDR*.rs12502008[TG]	(0[AA],0[TT],1[TG])(0[AA],1[CT],0[GG])(0[AA],2[CC],0[GG])(0[AA],2[CC],1[TG])(1[GA],0[TT],0[GG])(1[GA],1[CT],1[TG])(1[GA],2[CC],2[TT])(2[GG],0[TT],1[TG])(2[GG],1[CT],1[TG])(2[GG],1[CT],2[TT])	0.038	4.028E-06	5.418	2.642-11.114
VEGFR1	*MICAL2*.rs1564947[AG] *MICAL2*.rs954428[GA] *NEDD4*.rs12232351[AT]	(0[GG],0[AA],0[TT])(0[GG],0[AA],2[AA])(0[GG],1[GA],0[TT])(0[GG],1[GA],1[AT])(0[GG	<0.001	3.505E-08	14.855	5.693-38.761

		],2[GG],0[TT])(1[AG],0[AA],0[TT])(1[AG],1[GA],1[AT])(2[AA],2[GG],0[TT])(2[AA],2[GG],2[AA])				
VEGFR2	*DNM2*.rs7246673[TG] *NRP1*.rs10827227[TC] *SCUBE2*.rs7106593[GT]	(1[TG],0[CC],1[GT])(1[TG],1[TC],0[TT])(1[TG],1[TC],2[GG])(1[TG],2[TT],0[TT])(1[TG],2[TT],1[GT])(1[TG],2[TT],2[GG])(2[TT],0[CC],2[GG])(2[TT],1[TC],1[GT])(2[TT],1[TC],2[GG])(2[TT],2[TT],1[GT])	<0.001	7.062E-09	8.712	4.186-18.129
VEGFR3	*CHRM3*.rs1782357[TC] *CHRM3*.rs685960[CT] *TMEM52B*.rs10505752[TC]	(0[CC],0[TT],0[CC])(0[CC],1[CT],0[CC])(1[TC],0[TT],1[TC])(1[TC],0[TT],2[TT])(1[TC],1[CT],0[CC])(2[TT],0[TT],2[TT])(2[TT],1[CT],0[CC])(2[TT],1[CT],2[TT])	<0.001	5.721E-08	8.030	3.784-17.038
PIGF	*FLT1*.rs3936415[AG] *FLT1*.rs11149523[AG] *NRP1*.rs2073320[TC]	(0[GG],0[GG],0[CC])(0[GG],0[GG],1[TC])(0[GG],1[AG],0[CC])(0[GG],1[AG],2[TT])(0[GG],2[AA],0[CC])(0[GG],2[AA],2[TT])(1[AG],0[GG],2[TT])(1[AG],1[AG],1[TC])(1[AG],2[AA],1[TC])(1[AG],2[AA],2[TT])	<0.001	4.218E-07	12.996	4.812-35.103

		(2[AA],0[GG],0[CC])(2[AA],2[AA],0[CC])				

CI: confidence interval; OR: odds ratio; SNP: single nucleotide polymorphism.
0, 1, and 2 in the High Risk Genotype column refer to additive coding, where the number refers to the number of minor alleles in the genotype.
Square brackets in the Top Model SNPs column indicate major and minor alleles for each SNP; in which the first letter represents the minor allele and the second letter represents the major allele. The High risk genotypes column lists genotypes which were found by GMDR 0.9 to be high risk for poor survival. High-risk genotypes have the following format: the items between each pair of parentheses specify a genotype which is high risk for poor survival according to the GMDR output, presented in the order of the SNPs listed in the Top model SNP column. e.g. for top model SNPs FLT1.rs3936415[AG]_FLT1.rs11149523[AG]_NRP1.rs2073320[TC], genotypes (0[GG],0[GG],0[CC]), rs3936415 = GG, rs11149523 = GG, and rs2073320 = CC were classified as high risk by the GMDR 0.9 procedure. Whenever a SNP with a main effect was identified in 1-way analysis, the analysis was repeated with that SNP removed from the dataset (i.e. successive iterations). *FLT1* is also known as *VEGFR1*; *KDR* is also known as *VEGFR2*; *FLT4* is also known as *VEGFR3*; and *PGF* is also known as *PIGF*.

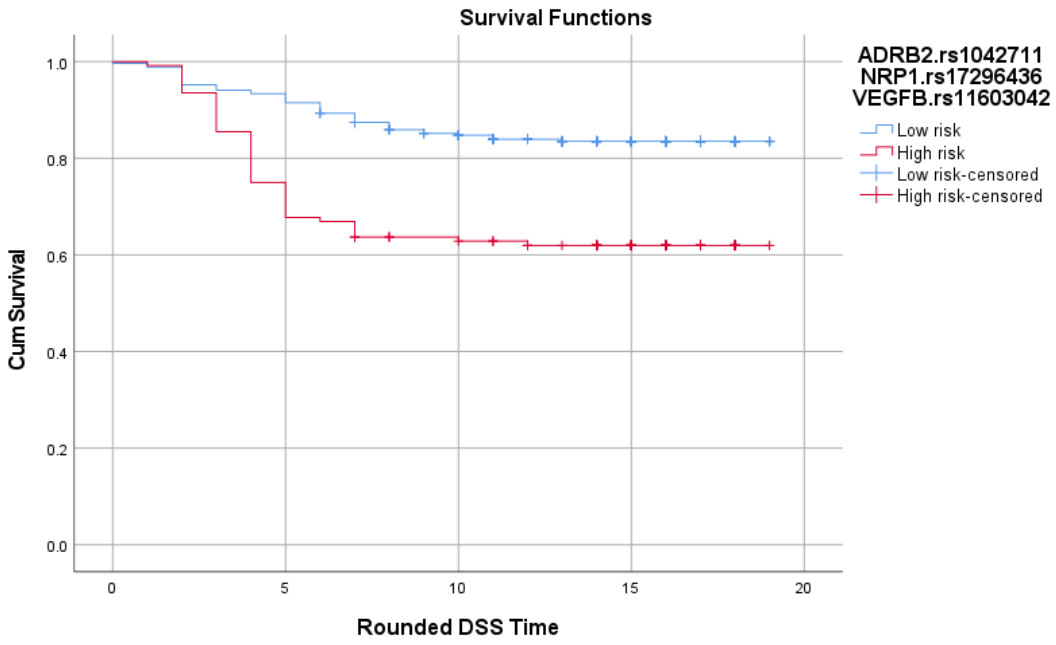


Figure 2.3: Kaplan-Meier curve for 3-way GMDR analysis, VEGFB dataset. *Log-rank p = 6.61897020900234E-07. Red: High risk genotypes: (TT,AA,TG), (TT,AA,TT), (TT,GA,TT), (TT,GG,TG), (CT,AA,TT), (CT,GA,GG), (CT,GG,GG), (CT,GG,TG), (CC,GA,GG), (CC,GA,TG), (CC,GA,TT). Blue: All others except (CT,GG,TT) and (CC,GG,TT). The vertical lines on the curves denote the censored patients (e.g. patients alive at the last follow up time). X and Y axis show the follow-up time (in years; rounded) and cumulative survival, respectively.*

2.5.1 Comparison of Cox-MDR and GMDR 0.9 results

Both Cox-MDR and GMDR 0.9 identified *RNF123*.rs11130216 SNP in the 1-way analysis of the *PIGF* network. In both cases, the same genotypes were identified as high-risk and were associated with DSS in multivariable models. All other significant interactions were identified by either of the programs. Our results, hence, showed that there was little overlap between the results provided by Cox-MDR and GMDR 0.9. This may be initially attributed to the use of different scoring systems and response variables by these programs. However, Cox-MDR was the software which identified the *MMP27*.rs11225388 variant, as well as the high-risk/low-genotype classification, that was previously identified to be associated with OS in a highly similar patient cohort[33]. Of note, this SNP had the strongest association in that dataset, so it is being identified by Cox-MDR and in all of the 20 1-way runs as the best SNP is quite striking (**Appendix 1-Table S6**). This SNP, however, was missed by GMDR 0.9. In addition, in GMDR 0.9, it was observed that there was no obvious way in which ties between "best models" (i.e. multiple "best models" with equal CVC values when selecting the best model) were being resolved. To test the effect of SNP order in the input data file, *MMP27*.rs11225388, a SNP with a known statistical association (see above), was moved to the beginning of the data file. This change resulted in significantly different GMDR 0.9 results (making rs11225388 the top SNP identified for this analysis) and thus, showed that input SNP order can affect results when the CVC is 1 or 2, out of a possible 5 (when multiple best models have the same CVC). Further observation confirmed that the earliest SNP in the dataset is chosen by GMDR 0.9 in the event of a CVC tie. Therefore, this not only

explains why GMDR 0.9 missed this SNP, but also an important limitation of this and any other MDR software that uses CVC to pick the best model. Despite its limitation, it is worth noting that GMDR 0.9 also identified a number of models that were missed by Cox-MDR and distinguished patients based on their significantly different outcome risks (**Table 2.2**; **Table 2.4**).

2.6 **Discussion**

In this study, we explored the functionality and feasibility of two MDR-based programs, Cox-MDR[105] and GMDR 0.9[101] and applied them to examine single-locus and multi-loci interactions in MMP family and VEGF interaction network genes in relation to survival outcome risks in colorectal cancer. Our results identified novel and statistically significant interactions that predicted the survival outcomes in colorectal cancer. Our results also showed that these two programs generally yielded different top MDR models and interactions, hence, they can be considered complementary while examining SNP interactions. To our knowledge, this is the first large-scale MDR analysis study that examined SNP interactions in relation to colorectal cancer outcomes.

Interactions among variables are understudied in cancer research. It is possible that the interactions among genetic variables, such as SNPs, play a role in survival outcomes biologically. Hence, limiting a study to associations of individual SNPs and survival outcomes has the potential to miss not only genetic relationships but also important biological information. In this regard, there has been little work done on studying multi-loci interactions in colorectal cancer with respect to survival outcomes, especially using a large number of variants. For example, limited MDR-based interaction analyses were

conducted[94–97,136,159], investigating the interactions among a small number of polymorphisms (n=5–17). These studies identified interacting polymorphisms that are associated with treatment response and/or survival outcomes. Therefore, while there has been little research on multi-loci interactions in colorectal cancer with respect to survival outcomes, there is also great potential in this area of research – this was our motivation to conduct this study. Additionally, in this study, we prioritized biologically relevant genes with well-known roles in disease progression in cancer: MMP family of genes and genes whose protein products were members of the protein interaction networks of seven separate VEGF-family proteins. Protein products of MMP family genes are involved in tissue remodeling, some of them have abnormalities associated with tumor invasion, tumor microenvironment, or metastasis[29]. VEGF family of proteins are also involved in important cellular processes, and include VEGF ligands and receptors with roles in angiogenesis or lymphangiogenesis – two cellular mechanisms involved in tumor growth, invasion, and metastasis[28,30,31]. Therefore, the results of this study have the potential to provide new insights into the relationship of these genes, molecular pathways, and processes with the outcome risk in colorectal cancer.

In this study, we first verified whether the MDR-based methods are indeed useful in distinguishing genotypes as high-risk and low risk. In the analysis of the interactions among the MMP family gene SNPs, 1-way Cox-MDR analysis was in fact able to identify a SNP in the dataset which has a known main effect, i.e. associated with the OS in the patient cohort under dominant genetic model (*MMP27*.rs11225388). This SNP was previously examined in our lab using a similar patient cohort and using Cox regression

method and it had the strongest association in the SNP set[33]. This previous study had also shown the dominant genetic model as the best model explaining the relationship of the genotypes of this SNP with patient overall survival times. In the current study, association of *MMP27*.rs11225388 under the dominant genetic model with the OS times in the study cohort was also confirmed by Cox-MDR, classifying the high-risk and low risk genotypes correctly (**Table 2.1**). Therefore, Cox-MDR was able to identify a SNP significantly associated with the outcome measure and its genetic model correctly, which increased our confidence in Cox-MDR results, though Cox-MDR did not identify any multi-loci interactions in this data set.

In contrast, GMDR 0.9 identified two novel multi-loci interactions in the MMP dataset; *MMP16*.rs7817382_*MMP24*.rs2254207 and *MMP16*.rs2664369_*MMP20*.rs11225332_*MMP2*.rs11639960 (**Table 2.2**). Interestingly, both of the variants identified in 2-way analysis (*MMP16*.rs7817382 and *MMP24*.rs2254207) are also eQTLs and associated with the expression levels of *MMP16* and *MMP24-AS1* genes, respectively (**Appendix 1-Table S7**). Protein products of *MMP16* and *MMP24* are known to interact physically with pro-MMP2 and activate it by means of proteolytic cleavage[160,161]. *MMP2* has been linked to several human cancers, including colorectal cancer previously[162–167]. Therefore, it is possible that the role of both *MMP16* and *MMP24* in affecting the action of *MMP2* could explain the biology underlying the interaction identified by 2-way GMDR analysis. Additionally, one of the SNPs identified in the 3-way GMDR 0.9 analysis, *MMP2*.rs11639960, is an eQTL, affecting the expression levels of the gene called *LPCAT2*. *LPCAT2* is known to affect

response to chemotherapy in colorectal cancer patients through an association with lipid droplet formation[168]. This SNP was also associated with prostate[169], and ovarian cancer risks[170], as well as overall survival in colorectal cancer[171]. Two of the genes identified in 3-way GMDR analysis are known to be associated with colorectal cancer. As mentioned above, *MMP2* has been shown to be overexpressed in colorectal cancer tumors compared to normal tissues[163,166], and is associated with metastatic tumor phenotype[163,166] and shorter survival times in colorectal cancer[163]. *MMP16* has a similar relationship to colorectal tumors[172]. *MMP20*, on the other hand, is a much less investigated member of the MMP family, but was found to be expressed in colorectal tumors in a study with small number of samples[173]. This 3-way interaction (*MMP16*.rs2664369_*MMP20*.rs11225332_*MMP2*.rs11639960) had a low p-value (1.3E-08) in the multivariable regression analysis and is, therefore, a particularly interesting example of both the potential biological roles of MMP gene variants in disease outcomes and the potential utility multi-loci interactions to help classifying patients based on their different outcome risks.

In the analyses of the seven VEGF interaction networks (*VEGFA*, *VEGFB*, *VEGFC*, *PIGF*, *VEGFR1*, *VEGFR2*, *VEGFR3* networks), similar to MMP gene analyses, MDR programs identified generally different results (e.g. interactions and SNPs). There is not any report linking the 1-way SNP identified by both programs with colorectal or other cancers (*RFN123*.rs11130216). However, both programs were again able to identify previously unknown and significant interactions. For example, the most significant interaction associated with disease-specific survival was detected in the 3-way Cox-MDR

analysis including the *CHRM3*.rs665159_*EPN1*.rs509955_*PTGER3*.rs1327460 variants (*VEGFR3* network; p=2.21E-09; **Table 2.3**). All of these genes were previously linked to cancer or tumor invasion. For example, high *CHRM3* levels are linked to invasion and metastasis in colon cancers[174,175]; loss of *EPN1* was linked to elevated *VEGFR2* degradation and disorganized angiogenesis[176]; and elevated *PTGER3* levels was linked to shorter survival times in cervical cancers[177]. On the other hand, the most significant GMDR 0.9 3-way model included variants from the *ADRB2*, *NRP1*, and *VEGFB* genes (logistic regression p-value=2.4E-09; **Table 2.4**). All three genes have been shown to be associated with colorectal cancer progression[178–180]. Also, while none of the variants identified in this study were missense or non-sense variants, according to GTEx[157] and RegulomeDB[155,156] a number of the SNPs identified were eQTLs (**Appendix 1 – Table S7**). Together with our results in the MMP gene analysis, the fact that the identified genes and/or interacting SNPs have been previously linked to colorectal cancer and/or tumor aggression, and in some cases, are associated with gene expression levels, make these multi-loci interactions highly promising candidates for future research.

We must also comment about the MDR-based programs that we utilized in this study. Cox-MDR and GMDR 0.9, while both have proven capable of finding significant models within the datasets (albeit often different models), they vary significantly in their functionality, operation, and resource usage. Cox-MDR was provided to us by the authors as a small collection of R functions, and as such did not have the full functionality we needed for our analyses, and therefore required further efforts to run. Many of these functions/features, on the other hand, were available in GMDR 0.9, such as returning

detailed outputs (including the output of high risk/low risk genotype information), the ability to set random seeds, and permutation testing. GMDR 0.9 is also readily available for download online. In contrast, an important feature possessed by Cox-MDR and missing from GMDR 0.9 is the ability to use testing balanced accuracy (TBA) score, as an alternative to CVC, to pick a best model from the cross-validation folds. GMDR 0.9 has a limitation that if two models tie for the best model among the cross-validation folds, then the model starting with the first SNP in the input dataset is chosen. This obviously has the potential to miss significant models as equally high-scoring models will be silently ignored by the software. This is an issue when using CVC to pick a best model more so than TBA (an option available in Cox-MDR), as when CVC is low it is quite likely that two or more models will tie for best model (used in GMDR 0.9; as we discuss earlier, GMDR 0.9 has missed identifying *MMP27*-rs11225388 in its 1-way analysis because of how it selects the top models (i.e. CVC and the order of data in the input files). This is rarely an issue while using TBA (that can be used in Cox-MDR) for the same purpose because as a floating point number with much higher variability than CVC, a tie is unlikely. Therefore, Cox-MDR using the TBA option overall gives results with less random model selection than GMDR 0.9, and this is an important strength of Cox-MDR. Despite its limitations, GMDR 0.9 also identified interactions that were missed by Cox-MDR.

Additionally, both Cox-MDR and GMDR 0.9 proved to have different resource usage difficulties and requirements. The Cox-MDR software cannot examine interactions in parallel, and thus, is significantly slower than GMDR 0.9. Our VEGFR2 3-way

analysis of 747 SNPs took approximately 18 days to complete on the local computing cluster whereas on a similar dataset GMDR 0.9 took only 12 hours. GMDR 0.9, on the other hand, has extremely large memory requirements. For the largest of our aforementioned analyses, GMDR 0.9 required a massive 220 gigabytes of RAM to complete successfully, which at the time of writing is a very large amount for a researcher to be able to obtain even on a computing cluster. In comparison, Cox-MDR only required 15 gigabytes of RAM, practically obtainable on consumer hardware. An additional resource usage issue for GMDR 0.9 is that the permutation testing procedure is performed using a Perl script external to the Java binary which contains the main program. This script uses the user's hard drive as memory, greatly slowing down the permutation testing procedure. For a very high number of permutations this may become a significant issue. Overall, while MDR-based data reduction methods allow researchers to examine large number of interactions, in our experience, both programs have unique strengths, limitations, and feasibility concerns while examining large datasets. Therefore, while they can be considered complementary while examining SNP interactions, application of these programs widely will likely be dependent on further development.

One limitation of this study is that the patients included are all of Caucasian ancestry. We also limited our work to common SNPs and genes from autosomal chromosomes, therefore, the potential interactions among rare SNPs and MMP/VEGF-interactor genes located in X or Y chromosomes remain unexamined. Our results are exploratory, therefore replication studies are needed to confirm whether these SNPs/interactions have prognostic value in the clinic. The genes were limited to select

genes related to cancer and progression, therefore further studies are needed to examine the potential interactions in other genes/interaction networks. Our study also has several strengths. This is one of the first studies that applied MDR-based approaches while examining survival outcomes in colorectal cancer, and the first one, in our knowledge, that examined such relatively large number of interactions (~90 million). We explored and applied two different MDR-based programs, one using the survival times (Cox-MDR) and the other 5-year survival status (GMDR 0.9) with a slightly different methodology that allowed us to comprehensively examine the interactions and compare the programs' utility. The patient cohort is a well annotated cohort. Additionally, the use of cross-validation and permutation testing, as well as the repeating the Cox-MDR/GMDR 0.9 runs (20 times) to identify the most consistent best models (called top models in this study) were critical and helped reduce the false-positive findings. More importantly, our results demonstrated that MDR can be powerful in detecting interactions among genetic variants in prognostic studies and the novel 2-way and 3-way SNP interactions identified in this study bring a new depth to colorectal cancer and prognostic research.

In conclusion, we performed a two-part study applying two MDR-based programs to examine the SNP interactions in relation to patient outcomes in colorectal cancer. Our work indicates that MDR-based programs can be quite useful in examining the interactions among the genotypes/SNPs while examining the novel prognostic markers in colorectal cancer. Our results also suggest the presence of novel SNPs and interactions in

MMP and VEGF family genes that are associated with the patient outcomes in colorectal cancer. These SNPs are excellent candidates for further biomarker studies.

2.7 **Data availability statement**

The datasets presented in this article are not readily available. Data that support the findings of this study are available from the Newfoundland Colorectal Cancer Registry/Memorial University. However, restrictions apply to the availability of this data, and so data are not publicly available. The data used in this study cannot be made publicly available as patients were not consented to make their data publicly available or accessible. Clinical and genetic data are available from the Newfoundland Colorectal Cancer Registry (NFCCR) upon reasonable request for researchers who meet the criteria for access to confidential data. Request to access the datasets should be directed to Newfoundland Colorectal Cancer Registry (PP; pparfrey@mun.ca) and Research, Grant, and Contract Services (rgcs@mun.ca) at Memorial University of Newfoundland, St. John's, NL, Canada, and the ethics approval shall be obtained from the Health Research Ethics Board (HREB), Ethics Office, Health Research Ethics Authority, Suite 200, 95 Bonaventure Avenue, St. John's, NL, A1B 2X5, Canada. The Cox-MDR code can be requested from Dr. Seungyeoun Lee. The GMDR 0.9 code can be requested from the developers, Drs. Xiang-Yang Lou, Jun Zhu, or Ming D. Li.

2.8 Ethics statement

The studies involving human participants were reviewed and approved by the Health Research Ethics Authority of Newfoundland and Labrador (HREB). Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

2.9 Funding

This study was supported by the Memorial University Seed, Bridge, and Multidisciplinary research funds (Seed funds to SS).

2.10 Acknowledgements

Authors thank the patients recruited to and investigators/staff at Newfoundland Colorectal Cancer Registry (NFCCR); Dr. Seungyeoun Lee for allowing to use the Cox-MDR code; Dr. Guobo Chen for correspondence related to their GMDR 0.9 software; staff at the Provincial Tumor Registry-NL and NLCHI for their help with the clinical data; Lucas Gillingham from CHIA, who helped with computational issues, especially during the COVID-19 pandemic/lock-down. This study was supported by the Memorial University Seed, Bridge, and Multidisciplinary research funds (Seed funds to SS). SS is a senior scientist of Beatrice Hunter Cancer Research Institute (BHCRI).

2.11 **Conflict of interest statement**

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Chapter 3: Discussion and conclusions

Colorectal cancer affects many people worldwide and in Canada. The highest incidence and mortality for this disease, among all Canadian provinces, is in Newfoundland and Labrador[2], making colorectal cancer both important and interesting to study in this province. My research described in this thesis aimed to identify new prognostic markers in colorectal cancer using the clinical and genetic data of a colorectal cancer cohort from Newfoundland and Labrador.

Identifying prognostic markers, including genetic markers, can support clinic management and survival outcomes of patients. While valuable research has been performed in colorectal cancer – including GWASs examining the association of genetic variants with clinical outcomes – these studies often look at only the associations of individual SNPs, neglecting potential SNP interactions[64,72,73,75,83–86].

To help fill this knowledge gap, my research included two specific aims. First, I compared two MDR-based programs, GMDR 0.9 and Cox-MDR using a patient dataset. GMDR 0.9 is a generalized version of the MDR algorithm, which can use a generalized linear model to evaluate patient risk[101]. This expands the number of possible study designs to which MDR can be applied and also allows for the inclusion of covariates in analysis. Cox-MDR[105] is a specification of GMDR, which utilizes the Cox regression method, and thus was able to more appropriately take advantage of the longitudinal survival data that was available to me. I applied both of these programs to data obtained from patients recruited to the Newfoundland Familial Colorectal Cancer Registry (NFCCR)[138], examining the relationship between MMP SNP interactions and overall

survival. My results demonstrated that neither of the algorithms was better than the other (i.e., they produced largely different models and hence complementary results). Therefore, I decided to continue with using both programs in my second aim. In the 2nd part of my study, I examined the relationship of a larger-scale SNP interaction set with disease specific survival by focusing on seven datasets (the protein-protein interaction networks of seven VEGF ligands and receptors) using both GMDR 0.9 and Cox-MDR. During these analyses, I chose to work on the MMP and VEGF gene family members, considering their established and interesting biological roles in tumor progression, metastasis, and patient outcomes[28–31].

Through my analyses, I found a set of significant and previously unknown interactions that separate patients based on their outcome risks. In the first part of my study, while studying SNP interactions in the MMP genes, I found three significant models. Cox-MDR found a significant 1-way model consisting of rs11225388, a SNP that was previously found to be associated with colorectal cancer survival in our lab[33]. Cox-MDR, however, did not find any significant multi-SNP models. Interestingly, GMDR 0.9 initially missed rs11225388, but identified significant 2-way and 3-way models that were completely missed by Cox-MDR. I later identified that GMDR 0.9 initially missed rs11225388 due to the use of CVC in selecting best models – by default this software selects the first SNP with the highest CVC in the dataset. As discussed in Chapter 2, this finding has important implications for researchers using CVC in selecting MDR models.

In the VEGF interactome part of my study, Cox-MDR identified a total of three significant models: a 1-way interaction in the *PIGF* interactome data set (which was also identified by GMDR 0.9. This interaction was the only one identified by both programs in this study), as well as significant 2-way and 3-way interactions in the *VEGFR3* data set. GMDR 0.9, on the other hand, found far more interactions (eleven 1-way interactions, six 2-way interactions, and seven 3-way interactions in the VEGF interactome data sets). These are novel interaction models that can predict prognosis in colorectal cancer. Overall, GMDR 0.9 identified a higher number of interaction models than Cox-MDR. In addition, except for a single model, all identified interaction models distinguished between patients with low and high outcome risk in multivariable regression models (i.e. associations of the high-risk/low-risk genotype categories identified by the MDR models with disease outcomes were verified by multivariable logistic regression, in the case of GMDR 0.9 models, or Cox regression models in the case of Cox-MDR models, when adjusted for clinical prognostic markers).

The literature shows that all 2- and 3-way, and vast majority of the 1-way interactions that I identified are novel. With the exception of rs11225388, none of the variants that I identified in my work were previously found to be associated with cancer outcomes, although several of the genes had been previously implicated. For example, *CHRM3*, *EPN1*, and *PTGER3* from a 3-way interaction found in my *VEGFR3* Cox-MDR interactome analysis were found to be related to metastasis in colorectal cancer[174], abnormal angiogenesis[176], and shorter survival times in cervical cancers[177], respectively. Similarly, *ADRB2*, *NRP1*, and *VEGFB* from my 3-way *VEGFB* interactome analysis had

previously been associated with colorectal cancer progression[178–180]. All of the variants that I found were in non-coding regions, and thus their functions are likely to be regulatory in nature. As also discussed in Chapter 2, our bioinformatics analyses revealed that several of the SNPs were known to be eQTL loci, which associate with gene expression regulation. For example, rs11639960 of the *MMP2* gene was an eQTL known to correlate with expression levels of *LPCAT2*, a gene that had been previously associated with chemotherapy response for colorectal cancer[181]. This may be a clue to understanding the biology underlying the statistical effect which I observed. While replication of my results in other patient cohorts is needed prior to any clinic usage, further studies on these variants and interactions may generate new biological knowledge on colorectal cancer prognosis, and as such, have the potential to be of great value.

Overall, my comparison between the results by GMDR 0.9 and Cox-MDR revealed that both programs gave different and complementary results. Both programs were also found to have different computer memory and time requirements and suffered from inefficiencies in resource usage, in both computer memory and time for GMDR 0.9 and Cox-MDR, respectively. This may be in part due to the additions I made to the software (e.g., to obtain intermediary results for models, or saving analysis files for potential future use). The largest single data set that I analyzed had a total of 747 SNPs. Since the complexity of combinatorial problems grows very quickly as the number of SNPs in a data set increases, this data set constitutes approximately 70,000,000 combinations that had to be explored in a single MDR analysis. This dataset took approximately 220GB of Random Access Memory (RAM) to analyze, per 3-way run,

using GMDR 0.9. In contrast, 17 days were required to run 3-way analyses with Cox-MDR for this data set. Cox-MDR's RAM usage, and GMDR 0.9's run-time, on the other hand, were very reasonable (15GB of RAM and 12 hours, respectively). This exposes clear limits on the utility of these two programs and the size of the analyses that they can currently perform. For example, consider an analysis of larger datasets in the future, which is one of our lab's goals. A modest genome-wide analysis on 50,000 SNPs would result in a space of 2.08×10^{13} 3-way combinations to explore. This is a factor of nearly 300,000 times larger than my largest analysis on the VEGF interactome datasets and would be clearly impractical using the software that I used. An MDR-based program that was written to be both memory and time efficient would allow significantly larger analyses to be performed, but even without the inefficiencies of the software I used, there are limits on the capabilities of MDR algorithms, which exhaustively explore all possible combinations of SNPs. As such, my conclusion is that somewhat larger data sets will require MDR-based methods different from the those which I used, and to examine much larger datasets (e.g., genome-wide), algorithms that exhaustively search all possible combinations may need to be replaced with other methods, such as those that narrow-down the number of SNPs to be examined. There are several examples of software that was designed to do just that, for example, one method which filters out SNPs before interaction analysis via a "group-sampling" method[182] and SNPHarvester[183] which filters SNPs via their relationship to the disease of interest before searching the interaction space.

The total number of possible interactions examined in my study is around 90 million. This is the largest interaction analysis ever conducted in colorectal cancer prognosis. Before my work, the largest study published in colorectal cancer examining interactions in relation to prognosis only looked at 17 SNPs, as compared to 1,517 SNPs, which I examined in the VEGF part of my study. My research is therefore currently the most comprehensive SNP interaction analysis in colorectal cancer prognosis. I am also glad that I have found novel candidate interactions associated with colorectal cancer outcomes and that I have published these and many interesting results in my manuscript[128]. These achievements were very rewarding. I am confident that my work not only contributes to the field of colorectal cancer research, but that also to biomarker discovery and interactions. I hope that my efforts, work, and findings will also inspire others to explore SNP interactions in colorectal cancer and other human diseases, an area of human genetics research that is both under-studied and full of potential.

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018).
2. Government of Canada. Canadian cancer statistics 2021. at https://www.canada.ca/content/dam/phac-aspc/documents/services/reports-publications/health-promotion-chronic-disease-prevention-canada-research-policy-practice/vol-41-no-11-2021/canadian-cancer-statistics-2021.pdf (2021).
3. Brenner, H., Kloor, M. & Pox, C. P. Colorectal cancer. *Lancet Lond. Engl.* **383**, 1490–1502 (2014).
4. Burt, R. W. *et al.* Colorectal cancer screening: clinical practice guidelines in oncology. *JNCCN J. Natl. Compr. Cancer Netw.* **11**, 1538–1575 (2013).
5. Lee, S. Risk factors for colorectal cancer. *Canadian Cancer Society* https://cancer.ca/en/cancer-information/cancer-types/colorectal/risks.
6. Kanth, P., Grimmett, J., Champine, M., Burt, R. & Samadder, J. N. Hereditary colorectal polyposis and cancer syndromes: a primer on diagnosis and management. *Off. J. Am. Coll. Gastroenterol. ACG* **112**, 1509–1525 (2017).
7. Boland, P. M., Yurgelun, M. B. & Boland, C. R. Recent progress in Lynch syndrome and other familial colorectal cancer syndromes. *CA. Cancer J. Clin.* **68**, 217–231 (2018).
8. Cerretelli, G., Ager, A., Arends, M. J. & Frayling, I. M. Molecular pathology of Lynch syndrome. *J. Pathol.* **250**, 518–531 (2020).

9. Ma, H. *et al.* Pathology and genetics of hereditary colorectal cancer. *Pathology (Phila.)* **50**, 49–59 (2018).

10. Kerr, S. E., Thomas, C. B., Thibodeau, S. N., Ferber, M. J. & Halling, K. C. APC germline mutations in individuals being evaluated for familial adenomatous polyposis: a review of the Mayo Clinic experience with 1591 consecutive tests. *J. Mol. Diagn.* **15**, 31–43 (2013).

11. Cheadle, J. P. & Sampson, J. R. MUTYH-associated polyposis—from defect in base excision repair to clinical genetic testing. *DNA Repair* **6**, 274–279 (2007).

12. Stanich, P. P. & Pearlman, R. Hereditary or not? Understanding serrated polyposis syndrome. *Curr. Treat. Options Gastroenterol.* **17**, 692–701 (2019).

13. Yamagishi, H., Kuroda, H., Imai, Y. & Hiraishi, H. Molecular pathogenesis of sporadic colorectal cancers. *Chin. J. Cancer* **35**, 4 (2016).

14. Thean, L. F. *et al.* Genome-wide association study identified copy number variants associated with sporadic colorectal cancer risk. *J. Med. Genet.* **55**, 181–188 (2018).

15. Lu, Y. *et al.* Large-scale genome-wide association study of East Asians identifies loci associated with risk for colorectal cancer. *Gastroenterology* **156**, 1455–1466 (2019).

16. Schmit, S. L. *et al.* Genome-wide association study of colorectal cancer in Hispanics. *Carcinogenesis* **37**, 547–556 (2016).

17. Mármol, I., Sánchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E. & Rodriguez Yoldi, M. J. Colorectal carcinoma: a general overview and future perspectives in colorectal cancer. *Int. J. Mol. Sci.* **18**, E197 (2017).

18. Thomas, M. *et al.* Genome-wide modeling of polygenic risk score in colorectal cancer risk. *Am. J. Hum. Genet.* **107**, 432–444 (2020).

19. Goel, A. & Boland, C. R. Recent insights into the pathogenesis of colorectal cancer. *Curr. Opin. Gastroenterol.* **26**, 47–52 (2010).

20. Huang, D. *et al.* Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer Metastasis Rev.* **37**, (2018).

21. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).

22. Bae, J. M., Kim, J. H. & Kang, G. H. Molecular subtypes of colorectal cancer and their clinicopathologic features, with an emphasis on the serrated neoplasia pathway. *Arch. Pathol. Lab. Med.* **140**, 406–412 (2016).

23. Popat, S., Hubner, R. & Houlston, R. S. Systematic review of microsatellite instability and colorectal cancer prognosis. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23**, 609–618 (2005).

24. Heath, A. P. *et al.* The NCI Genomic Data Commons. *Nat. Genet.* **53**, 257–262 (2021).

25. Lee, S. Staging cancer. *Canadian Cancer Society* https://cancer.ca/en/cancer-information/what-is-cancer/stage-and-grade/staging.

26. Colorectal Cancer Stages | Colorectal Cancer Alliance. https://www.ccalliance.org/colorectal-cancer-information/stage-of-diagnosis.

27. Miller, K. D. *et al.* Cancer treatment and survivorship statistics, 2019. *CA. Cancer J. Clin.* **69**, 363–385 (2019).

28. Hicklin, D. J. & Ellis, L. M. Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **23**, 1011–1027 (2005).

29. Hua, H., Li, M., Luo, T., Yin, Y. & Jiang, Y. Matrix metalloproteinases in tumorigenesis: an evolving paradigm. *Cell. Mol. Life Sci. CMLS* **68**, 3853–3868 (2011).

30. Lohela, M., Bry, M., Tammela, T. & Alitalo, K. VEGFs and receptors involved in angiogenesis versus lymphangiogenesis. *Curr. Opin. Cell Biol.* **21**, 154–165 (2009).

31. Alitalo, A. & Detmar, M. Interaction of tumor cells and lymphatic vessels in cancer progression. *Oncogene* **31**, 4499–4508 (2012).

32. Genentech. Avastin prescribing information. https://www.accessdata.fda.gov/drugsatfda_docs/label/2009/125085s0169lbl.pdf.

33. Dan, L. A. *et al.* No associations of a set of SNPs in the Vascular Endothelial Growth Factor (VEGF) and Matrix Metalloproteinase (MMP) genes with survival of colorectal cancer patients. *Cancer Med.* **5**, 2221–2231 (2016).

34. Levin, T. R. *et al.* Effects of organized colorectal cancer screening on cancer incidence and mortality in a large, community-based population. *Gastroenterology* **155**, 1383-1391.e5 (2018).

35. Eastern Health. Colon cancer screening program – cancer care. https://cancercare.easternhealth.ca/prevention-and-screening/colon-cancer-screening/ (2022).

36. Guba, M., Seeliger, H., Kleespies, A., Jauch, K.-W. & Bruns, C. Vascular endothelial growth factor in colorectal cancer. *Int. J. Colorectal Dis.* **19**, 510–517 (2004).

37. Kishore, C. & Bhadra, P. Current advancements and future perspectives of immunotherapy in colorectal cancer research. *Eur. J. Pharmacol.* **893**, 173819 (2021).

38. Cancer Research UK. Worldwide cancer statistics. https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer (2015).

39. van der Stok, E. P., Spaander, M. C. W., Grünhagen, D. J., Verhoef, C. & Kuipers, E. J. Surveillance after curative treatment for colorectal cancer. *Nat. Rev. Clin. Oncol.* **14**, 297–315 (2017).

40. Hansdotter, P. *et al.* Patterns and resectability of colorectal cancer recurrences: outcome study within the COLOFOL trial. *BJS Open* **5**, zrab067 (2021).

41. Galandiuk, S. *et al.* Patterns of recurrence after curative resection of carcinoma of the colon and rectum. *Surg. Gynecol. Obstet.* **174**, 27–32 (1992).

42. Lech, G., Słotwiński, R., Słodkowski, M. & Krasnodębski, I. W. Colorectal cancer tumour markers and biomarkers: Recent therapeutic advances. *World J. Gastroenterol.* **22**, 1745–1755 (2016).

43. Moghimi-Dehkordi, B. & Safaee, A. An overview of colorectal cancer survival rates and prognosis in Asia. *World J. Gastrointest. Oncol.* **4**, 71–75 (2012).

44. Yu, Y. *et al.* The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. *BMC Med.* **17**, 150 (2019).

45. Brenner, D. R. *et al.* Projected estimates of cancer in Canada in 2022. *CMAJ* **194**, E601–E607 (2022).

46. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

47. Nurk, S. *et al.* The complete sequence of a human genome. *Science* **376**, 44–53 (2022).

48. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

49. Karczewski, K. J. *et al.* The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res.* **45**, D840–D845 (2017).

50. ExAC project pins down rare gene variants. *Nature* **536**, 249–249 (2016).

51. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

52. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

53. Cannell, I. G., Kong, Y. W. & Bushell, M. How do microRNAs regulate gene expression? *Biochem. Soc. Trans.* **36**, 1224–1231 (2008).

54. Chen, S. & Shen, X. Long noncoding RNAs: functions and mechanisms in colon cancer. *Mol. Cancer* **19**, 167 (2020).

55. Ye, J.-J. & Cao, J. MicroRNAs in colorectal cancer as markers and targets: Recent advances. *World J. Gastroenterol.* **20**, 4288–4299 (2014).

56. Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693–705 (2007).

57. F. Zhao, R. ENCODE: deciphering function in the human genome. *Genome.gov* https://www.genome.gov/27551473/genome-advance-of-the-month-encode-deciphering-function-in-the-human-genome.

58. Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).

59. Chen, C.-Y., Chang, I.-S., Hsiung, C. A. & Wasserman, W. W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med. Genomics* **7**, 34 (2014).

60. Brookes, A. J. The essence of SNPs. *Gene* **234**, 177–186 (1999).

61. Gu, W. & Lupski, J. R. CNV and nervous system diseases--what's new? *Cytogenet. Genome Res.* **123**, 54–64 (2008).

62. Sønderby, I. E. *et al.* Effects of copy number variations on brain structure and risk for psychiatric illness: Large-scale studies from the ENIGMA working groups on CNVs. *Hum. Brain Mapp.* **43**, 300–328 (2022).

63. Werdyani, S. *et al.* Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. *Cancer Med.* **6**, 1220–1232 (2017).

64. Yu, Y. *et al.* A comprehensive analysis of SNPs and CNVs identifies novel markers associated with disease outcomes in colorectal cancer. *Mol. Oncol.* **15**, 3329–3347 (2021).

65. Garziera, M. *et al.* HLA-G 3'UTR polymorphisms impact the prognosis of stage II-III CRC patients in fluoropyrimidine-based treatment. *PLoS ONE* **10**, e0144000 (2015).

66. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

67. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426**, 789–796 (2003).

68. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostat. Oxf. Engl.* **13**, 762–775 (2012).

69. Savas, S. & Younghusband, H. B. dbCPCO: a database of genetic markers tested for their predictive and prognostic value in colorectal cancer. *Hum. Mutat.* **31**, 901–907 (2010).

70. Savas, S. & Liu, G. Genetic variations as cancer prognostic markers: review and update. *Hum. Mutat.* **30**, 1369–1377 (2009).

71. Negandhi, A. A. *et al.* MTHFR Glu429Ala and ERCC5 His46His polymorphisms are associated with prognosis in colorectal cancer patients: analysis of two independent cohorts from Newfoundland. *PLoS One* **8**, e61469 (2013).

72. Labadie, J. D. *et al.* Genome-wide association study identifies tumor anatomical site-specific risk variants for colorectal cancer survival. *Sci. Rep.* **12**, 127 (2022).

73. Penney, M. E., Parfrey, P. S., Savas, S. & Yilmaz, Y. E. A genome-wide association study identifies single nucleotide polymorphisms associated with time-to-metastasis in colorectal cancer. *BMC Cancer* **19**, 133 (2019).

74. Meng, Y. *et al.* Genome-wide association analyses identify CATSPERE as a mediator of colorectal cancer susceptibility and progression. *Cancer Res.* **82**, 986–997 (2022).

75. Pander, J. *et al.* Genome wide association study for predictors of progression free survival in patients on Capecitabine, Oxaliplatin, Bevacizumab and Cetuximab in first-line therapy of metastatic colorectal cancer. *PLoS One* **10**, e0131091 (2015).

76. Mimori, K., Tanaka, F., Shibata, K. & Mori, M. Review: single nucleotide polymorphisms associated with the oncogenesis of colorectal cancer. *Surg. Today* **42**, 215–219 (2012).

77. Radanova, M. *et al.* Single nucleotide polymorphisms in microRNA genes and colorectal cancer risk and prognosis. *Biomedicines* **10**, 156 (2022).

78. Naccarati, A. *et al.* Mutations and polymorphisms in TP53 gene--an overview on the role in colorectal cancer. *Mutagenesis* **27**, 211–218 (2012).

79. Zhang, K., Civan, J., Mukherjee, S., Patel, F. & Yang, H. Genetic variations in colorectal cancer risk and clinical outcome. *World J. Gastroenterol.* **20**, 4167–4177 (2014).

80. Phipps, A. I. *et al.* Association between colorectal cancer susceptibility loci and survival time after diagnosis with colorectal cancer. *Gastroenterology* **143**, 51–4.e4 (2012).

81. Xing, J. *et al.* GWAS-identified colorectal cancer susceptibility locus associates with disease prognosis. *Eur. J. Cancer* **47**, 1699–1707 (2011).

82. Abulí, A. *et al.* Genetic susceptibility variants associated with colorectal cancer prognosis. *Carcinogenesis* **34**, 2286–2291 (2013).

83. Xu, W. *et al.* A genome wide association study on Newfoundland colorectal cancer patients' survival outcomes. *Biomark. Res.* **3**, 6 (2015).

84. Penney, K. L. *et al.* Genetic variant associated with survival of patients with stage II-III colon cancer. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **18**, 2717-2723.e3 (2019).

85. Phipps, A. I. *et al.* Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis. *Carcinogenesis* **37**, 87–95 (2016).

86. Innocenti, F. *et al.* Genomic analysis of germline variation associated with survival of patients with colorectal cancer treated with chemotherapy plus biologics in CALGB/SWOG 80405 (Alliance). *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **27**, 267–275 (2021).

87. Li, P., Guo, M., Wang, C., Liu, X. & Zou, Q. An overview of SNP interactions in genome-wide association studies. *Brief. Funct. Genomics* **14**, 143–155 (2015).

88. Domingo, J., Baeza-Centurion, P. & Lehner, B. The causes and consequences of genetic interactions (epistasis). *Annu. Rev. Genomics Hum. Genet.* **20**, 433–460 (2019).

89. Stern, D. B., Anderson, N. W., Diaz, J. A. & Lee, C. E. Genome-wide signatures of synergistic epistasis during parallel adaptation in a Baltic Sea copepod. *Nat. Commun.* **13**, 4024 (2022).

90. Davis, B. H., Poon, A. F. Y. & Whitlock, M. C. Compensatory mutations are repeatable and clustered within proteins. *Proc. R. Soc. B Biol. Sci.* **276**, 1823–1827 (2009).

91. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109**, 1193–1198 (2012).

92. Song, Y. S., Wang, F. & Slatkin, M. General epistatic models of the risk of complex diseases. *Genetics* **186**, 1467–1473 (2010).

93. Hemani, G., Knott, S. & Haley, C. An evolutionary perspective on epistasis and the missing heritability. *PLoS Genet.* **9**, e1003295 (2013).

94. Afzal, S. *et al.* The association of polymorphisms in 5-fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. *Pharmacogenomics* **12**, 1257–1267 (2011).

95. Pander, J. *et al.* Pharmacogenetic interaction analysis for the efficacy of systemic treatment in metastatic colorectal cancer. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **22**, 1147–1153 (2011).

96. Sarac, S. B. *et al.* Data-driven assessment of the association of polymorphisms in 5-Fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. *Basic Clin. Pharmacol. Toxicol.* **111**, 189–197 (2012).

97. Hu, X. *et al.* Polymorphisms in DNA repair pathway genes and ABCG2 gene in advanced colorectal cancer: correlation with tumor characteristics and clinical outcome in oxaliplatin-based chemotherapy. *Cancer Manag. Res.* **11**, 285–297 (2018).

98. Jung, S. Y. *et al.* Pro-inflammatory cytokine polymorphisms in ONECUT2 and HNF4A and primary colorectal carcinoma: a post genome-wide gene-lifestyle interaction study. *Am. J. Cancer Res.* **10**, 2955–2976 (2020).

99. Ritchie, M. D. *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69**, 138–147 (2001).

100. Motsinger, A. A. & Ritchie, M. D. Multifactor dimensionality reduction: an analysis strategy for modelling and detecting gene-gene interactions in human genetics and pharmacogenomics studies. *Hum. Genomics* **2**, 318–328 (2006).

101. Lou, X.-Y. *et al.* A generalized combinatorial approach for detecting gene-by-gene and gene-by-environment interactions with application to nicotine dependence. *Am. J. Hum. Genet.* **80**, 1125–1137 (2007).

102. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinforma. Oxf. Engl.* **22**, 2173–2174 (2006).

103. Karayianni, K. N., Grimaldi, K. A., Nikita, K. S. & Valavanis, I. K. Mining nutrigenetics patterns related to obesity: use of parallel multifactor dimensionality reduction. *Int. J. Bioinforma. Res. Appl.* **11**, 233–246 (2015).

104. Sinnott-Armstrong, N. A., Greene, C. S., Cancare, F. & Moore, J. H. Accelerating epistasis analysis in human genetics with consumer graphics hardware. *BMC Res. Notes* **2**, 149 (2009).

105. Lee, S., Kwon, M.-S., Oh, J. M. & Park, T. Gene-gene interaction analysis for the survival phenotype based on the Cox model. *Bioinforma. Oxf. Engl.* **28**, i582–i588 (2012).

106. Greener, J. G., Kandathil, S. M., Moffat, L. & Jones, D. T. A guide to machine learning for biologists. *Nat. Rev. Mol. Cell Biol.* **23**, 40–55 (2022).

107. Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W. & O'Sullivan, J. M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinforma.* **2**, 927312 (2022).

108. Chen, J. W. & Dhahbi, J. Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods. *Sci. Rep.* **11**, 13323 (2021).

109. Feng, C. H., Disis, M. L., Cheng, C. & Zhang, L. Multimetric feature selection for analyzing multicategory outcomes of colorectal cancer: random forest and multinomial

logistic regression models. *Lab. Investig. J. Tech. Methods Pathol.* **102**, 236–244 (2022).

110. Al-Rajab, M., Lu, J. & Xu, Q. Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis. *Comput. Methods Programs Biomed.* **146**, 11–24 (2017).

111. Uppu, S., Krishna, A. & Gopalan, R. P. A review on methods for detecting SNP interactions in high-dimensional genomic data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **15**, 599–612 (2018).

112. Zhang, X., Zou, F. & Wang, W. Fastanova: an efficient algorithm for genome-wide association study. in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* 821–829 (Association for Computing Machinery, 2008). doi:10.1145/1401890.1401988.

113. Lunetta, K. L., Hayward, L. B., Segal, J. & Van Eerdewegh, P. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genet.* **5**, 32 (2004).

114. McKinney, B. A., Reif, D. M., Ritchie, M. D. & Moore, J. H. Machine learning for detecting gene-gene interactions. *Appl. Bioinformatics* **5**, 77–88 (2006).

115. Jiang, R., Tang, W., Wu, X. & Fu, W. A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinformatics* **10**, S65 (2009).

116. Tomita, Y. *et al.* Artificial neural network approach for selection of susceptible single nucleotide polymorphisms and construction of prediction model on childhood allergic asthma. *BMC Bioinformatics* **5**, 120 (2004).

117. Rekaya, R. & Robbins, K. Ant colony algorithm for analysis of gene interaction in high-dimensional association data. *R. Bras. Zootec.* **38**, (2009).

118. Zhang, Y. & Liu, J. S. Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* **39**, 1167–1173 (2007).

119. Han, B., Chen, X. & Talebizadeh, Z. FEPI-MB: identifying SNPs-disease association using a Markov Blanket-based approach. *BMC Bioinformatics* **12**, S3 (2011).

120. Han, B. & Chen, X. bNEAT: a Bayesian network method for detecting epistatic interactions in genome-wide association studies. *BMC Genomics* **12 Suppl 2**, S9 (2011).

121. Velez, D. R. *et al.* A balanced accuracy function for epistasis modeling in imbalanced datasets using multifactor dimensionality reduction. *Genet. Epidemiol.* **31**, 306–315 (2007).

122. Wang, B., Wang, M., Li, X., Yang, M. & Liu, L. Variations in the Wnt/β-Catenin pathway key genes as predictors of cervical cancer susceptibility. *Pharmacogenomics Pers. Med.* **13**, 157–165 (2020).

123. Yu, Y. *et al.* Polymorphisms of inflammation-related genes and colorectal cancer risk: a population-based case–control study in China. *Int. J. Immunogenet.* **41**, 289–297 (2014).

124. Fu, D. *et al.* Impact of vascular endothelial growth factor gene-gene and gene-smoking interaction and haplotype combination on bladder cancer risk in Chinese population. *Oncotarget* **8**, 22927–22935 (2017).

125. Yadav, A. *et al.* Association of Wnt signaling pathway genetic variants in gallbladder cancer susceptibility and survival. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **37**, 8083–8095 (2016).

126. Yu, Y. *et al.* Association of genetic variants in tachykinins pathway genes with colorectal cancer risk. *Int. J. Colorectal Dis.* **27**, 1429–1436 (2012).

127. Agarwal, G., Tulsyan, S., Lal, P. & Mittal, B. Generalized Multifactor Dimensionality Reduction (GMDR) analysis of drug-metabolizing enzyme-encoding gene polymorphisms may predict treatment outcomes in Indian breast cancer patients. *World J. Surg.* **40**, 1600–1610 (2016).

128. Curtis, A. *et al.* Examining SNP-SNP interactions and risk of clinical outcomes in colorectal cancer using multifactor dimensionality reduction based methods. *Front. Genet.* **13**, 902217 (2022).

129. Compton, C. C. *et al.* Prognostic factors in colorectal cancer. College of American Pathologists Consensus Statement 1999. *Arch. Pathol. Lab. Med.* **124**, 979–994 (2000).

130. Berian, J. R., Benson, A. B. & Nelson, H. Young age and aggressive treatment in colon cancer. *JAMA* **314**, 613–614 (2015).

131. Steele, C. W., Whittle, T. & Smith, J. J. Review: KRAS mutations are influential in driving hepatic metastases and predicting outcome in colorectal cancer. *Chin. Clin. Oncol.* **8**, 53 (2019).

132. Coleman, M. P. *et al.* Cancer survival in five continents: a worldwide population-based study (CONCORD). *Lancet Oncol.* **9**, 730–756 (2008).

133. Pathy, S., Lambert, R., Sauvaget, C. & Sankaranarayanan, R. The incidence and survival rates of colorectal cancer in India remain low compared with rising rates in East Asia. *Dis. Colon Rectum* **55**, 900–906 (2012).

134. Arnold, M. *et al.* Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**, 683–691 (2017).

135. Ziv, E. *et al.* Genome-wide association study identifies variants at 16p13 associated with survival in multiple myeloma patients. *Nat. Commun.* **6**, 7539 (2015).

136. Iglesias, D. *et al.* Effect of COX2 -765G>C and c.3618A>G polymorphisms on the risk and survival of sporadic colorectal cancer. *Cancer Causes Control* **20**, 1421–1429 (2009).

137. Green, R. C. *et al.* Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. *Fam. Cancer* **6**, 53–62 (2007).

138. Woods, M. O. *et al.* The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut* **59**, 1369–1377 (2010).

139. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

140. PLINK: Whole genome data analysis toolset. http://zzz.bwh.harvard.edu/plink/ (2017).

141. R Core Team. *R: a language and environment for statistical computing*. (R Foundation for Statistical Computing, 2017).

142. Motsinger, A. A. & Ritchie, M. D. The effect of reduction in cross-validation intervals on the performance of multifactor dimensionality reduction. *Genet. Epidemiol.* **30**, 546–555 (2006).

143. Edwards, T. L., Lewis, K., Velez, D. R., Dudek, S. & Ritchie, M. D. Exploring the performance of Multifactor Dimensionality Reduction in large scale SNP studies and in the presence of genetic heterogeneity among epistatic disease models. *Hum. Hered.* **67**, 183–192 (2009).

144. Gui, J. *et al.* A novel survival multifactor dimensionality reduction method for detecting gene-gene interactions with application to bladder cancer prognosis. *Hum. Genet.* **129**, 101–110 (2011).

145. De, R. *et al.* Identifying gene-gene interactions that are highly associated with Body Mass Index using Quantitative Multifactor Dimensionality Reduction (QMDR). *BioData Min.* **8**, 41 (2015).

146. Gola, D., Mahachie John, J. M., van Steen, K. & König, I. R. A roadmap to multifactor dimensionality reduction methods. *Brief. Bioinform.* **17**, 293–308 (2016).

147. *IBM SPSS Statistics for Windows*. (IBM Corp., 2017).

148. Stark, C. *et al.* BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535-539 (2006).

149. Oughtred, R. *et al.* The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci. Publ. Protein Soc.* **30**, 187–200 (2021).

150. BioGRID | Database of protein, chemical, and genetic interactions. https://thebiogrid.org/.

151. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).

152. Ensembl Genome Browser. http://grch37.ensembl.org/index.html.

153. Ensembl Archives. http://useast.ensembl.org/info/website/archives/index.html.

154. Savas, S. A curated database of genetic markers from the angiogenesis/VEGF pathway and their relation to clinical outcome in human cancers. *Acta Oncol. Stockh. Swed.* **51**, 243–246 (2012).

155. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

156. RegulomeDB. http://www.regulomedb.org/.

157. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

158. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).

159. Afzal, S. *et al.* Combinations of polymorphisms in genes involved in the 5-Fluorouracil metabolism pathway are associated with gastrointestinal toxicity in chemotherapy-treated colorectal cancer patients. *Clin. Cancer Res.* **17**, 3822–3829 (2011).

160. Llano, E. *et al.* Identification and characterization of human MT5-MMP, a new membrane-bound activator of progelatinase a overexpressed in brain tumors. *Cancer Res.* **59**, 2570–2576 (1999).

161. Zhao, H. *et al.* Differential inhibition of membrane type 3 (MT3)-matrix metalloproteinase (MMP) and MT1-MMP by tissue inhibitor of metalloproteinase (TIMP)-2 and TIMP-3 regulates pro-MMP-2 activation. *J. Biol. Chem.* **279**, 8592–8601 (2004).

162. van der Jagt, M. F. P., Wobbes, T., Strobbe, L. J. A., Sweep, F. C. G. J. & Span, P. N. Metalloproteinases and their regulators in colorectal cancer. *J. Surg. Oncol.* **101**, 259–269 (2010).

163. Dong, W. *et al.* Matrix metalloproteinase 2 promotes cell growth and invasion in colorectal cancer. *Acta Biochim. Biophys. Sin.* **43**, 840–848 (2011).

164. Wang, H.-L., Zhou, P.-Y., Zhang, Y. & Liu, P. Relationships between abnormal MMP2 expression and prognosis in gastric cancer: a meta-analysis of cohort studies. *Cancer Biother. Radiopharm.* **29**, 166–172 (2014).

165. Ren, F. *et al.* Overexpression of MMP family members functions as prognostic biomarker for breast cancer patients: a systematic review and meta-analysis. *PLoS One* **10**, e0135544 (2015).

166. Gao, M. *et al.* Expression analysis and clinical significance of eIF4E, VEGF-C, E-cadherin and MMP-2 in colorectal adenocarcinoma. *Oncotarget* **7**, 85502–85514 (2016).

167. Jia, H., Zhang, Q., Liu, F. & Zhou, D. Prognostic value of MMP-2 for patients with ovarian epithelial carcinoma: a systematic review and meta-analysis. *Arch. Gynecol. Obstet.* **295**, 689–696 (2017).

168. Cotte, A. K. *et al.* Lysophosphatidylcholine acyltransferase 2-mediated lipid droplet production supports colorectal cancer chemoresistance. *Nat. Commun.* **9**, 322 (2018).

169. Jacobs, E. J. *et al.* Polymorphisms in angiogenesis-related genes and prostate cancer. *Cancer Epidemiol. Prev. Biomark.* **17**, 972–977 (2008).

170. Velapasamy, S. *et al.* Influences of multiple genetic polymorphisms on ovarian cancer risk in Malaysia. *Genet. Test. Mol. Biomark.* **17**, 62–68 (2013).

171. Scherer, D. *et al.* Abstract 2188: Genetic variation in angiogenesis-related genes is associated with colorectal cancer risk and prognosis. *Cancer Res.* **74**, 2188–2188 (2014).

172. Wu, S., Ma, C., Shan, S., Zhou, L. & Li, W. High expression of matrix metalloproteinases 16 is associated with the aggressive malignant behavior and poor survival outcome in colorectal carcinoma. *Sci. Rep.* **7**, 46531 (2017).

173. Kraus, D., Reckenbeil, J., Perner, S., Winter, J. & Probstmeier, R. Expression pattern of Matrix Metalloproteinase 20 (MMP20) in human tumors. *Anticancer Res.* **36**, 2713–2718 (2016).

174. Cheng, K., Shang, A. C., Drachenberg, C. B., Zhan, M. & Raufman, J.-P. Differential expression of M3 muscarinic receptors in progressive colon neoplasia and metastasis. *Oncotarget* **8**, 21106–21114 (2017).

175. Felton, J., Hu, S. & Raufman, J.-P. Targeting M3 muscarinic receptors for colon cancer therapy. *Curr. Mol. Pharmacol.* **11**, 184–190 (2018).

176. Pasula, S. *et al.* Endothelial epsin deficiency decreases tumor growth by enhancing VEGF signaling. *J. Clin. Invest.* **122**, 4424–4438 (2012).

177. Heidegger, H. *et al.* The prostaglandin EP3 receptor is an independent negative prognostic factor for cervical cancer patients. *Int. J. Mol. Sci.* **18**, 1571 (2017).

178. Kamiya, T. *et al.* The preserved expression of neuropilin (NRP) 1 contributes to a better prognosis in colon cancer. *Oncol. Rep.* **15**, 369–373 (2006).

179. Jayasinghe, C., Simiantonaki, N. & Kirkpatrick, C. J. VEGF-B expression in colorectal carcinomas and its relevance for tumor progression. *Histol. Histopathol.* **28**, 647–653 (2013).

180. Ogawa, H. *et al.* Prognostic significance of β2-adrenergic receptor expression in patients with surgically resected colorectal cancer. *Int. J. Clin. Oncol.* **25**, 1137–1144 (2020).

181. Cotte, A. K., Aires, V., Ghiringhelli, F. & Delmas, D. LPCAT2 controls chemoresistance in colorectal cancer. *Mol. Cell. Oncol.* **5**, e1448245 (2018).

182. Prabhu, S. & Pe'er, I. Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease. *Genome Res.* **22**, 2230–2240 (2012).

183. Yang, C. *et al.* SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinforma.* **25**, 504–511 (2009).

184. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. The human genome browser at UCSC. *Genome Res* **12**, 996–1006 (2002).

185. Yates, B. *et al.* Genenames.org: the HGNC and VGNC resources in 2017. *Nucleic Acids Res.* **45**, D619–D625 (2017).

186. *Microsoft R Open*. (Microsoft Corporation, 2019).

187. RStudio Team. *RStudio: Integrated Development Environment for R*. (RStudio, Inc., 2015).

188. Génin, E. *et al.* Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe. *Orphanet J. Rare Dis.* **6**, 52 (2011).

189. Pearson, J. P. *et al.* Familial frontotemporal dementia with amyotrophic lateral sclerosis and a shared haplotype on chromosome 9p. *J. Neurol.* **258**, 647–655 (2011).

Appendices

Appendix 1: Supplementary material for Chapter 2

This Appendix contains the Supplementary Material published together with the manuscript described in Chapter 2 (Curtis et al. Frontiers in Genetics 2022 Aug 3;13:902217. doi: 10.3389/fgene.2022.902217).

It can be found online here: https://www.frontiersin.org/articles/10.3389/fgene.2022.902217/full#supplementary-material

Methods

Part 1. Analysis of MMP gene SNPs

Patient cohort

Table S1: Baseline characteristics of the 439 patients who are included in the Part 1 of the study.

MMP Project		
Variable	**N**	**%**

Age at Diagnosis	Median: 62; Range: 21 – 75 years	
Disease Stage		
I	72	16.40
II	174	39.64
III	146	33.26
IV	47	10.71
MSI Status		
Stable/MSI-low	389	88.61
MSI-high	50	11.39
Tumor Location		
Colon	298	67.88
Rectum	141	32.12
#OS Time	Median: 7; Range: 0 – 11 years	
#OS Status		
Alive	279	63.55
Dead	160	36.45
##5-Year OS Status		
Alive at 5 years	324	73.80
Dead at 5 years	115	26.20

MSI: Microsatellite instability; **OS:** Overall survival. #Used in Cox-MDR, Kaplan-Meier, and Cox regression analyses. ##Used in GMDR 0.9 and logistic regression analyses

Genes Selected for Part 1

Genomic locations for MMP genes (n=23) were identified using the UCSC genome browser[184]. When needed, alternate gene symbols were found using the HUGO gene nomenclature (HGNC) database[185]. Of the results returned upon searching for each MMP gene in the UCSC genome browser, the earliest start location among the UCSC gene locations listed was selected as the beginning of the genomic range and the latest end location among locations listed was chosen as the end of the genomic range (**Table S2**).

Table S2: MMP genes, their genomic locations, and the numbers of SNPs per gene that are included in Part 1 of this study

Gene	Chromosome	Genomic Range (bp)	Number of SNPs	SNPs
MMP1	11	102660641 - 102668966	7	rs2239008_A rs470558_A rs10488_A rs470215_G rs1938901_T rs7125062_C rs3213460_A

MMP2	16	555113081 - 55540586	16	rs2287074_A rs11639960_G rs1477017_G rs865094_G rs17301608_T rs1132896_C rs1053605_T rs866770_G rs9302671_T rs243845_T rs243843_G rs243842_C rs183112_A rs2287076_C rs243835_T rs10775332_T
MMP3	11	102706528 - 102714342	3	rs3025066_G rs3020919_T rs679620_A
MMP7	11	102391239 - 102401478	4	rs17886371_G rs14983_T rs2156528_A rs1996352_C
MMP8	11	102582526 - 102595685	8	kgp5394892_G rs1940475_C rs12365082_A rs7934972_A rs3740938_A rs2012390_C rs12803000_G rs2155052_C
MMP9	20	44637547 - 44645200	4	rs2274755_T rs17576_G rs2274756_A rs20544_C

MMP10	11	102641233 - 102651359	8	rs470168_A rs12290253_C rs547561_G rs12272341_A rs4431992_C rs2276108_G rs17860950_C rs486055_T
MMP11	22	24115036 - 24126503	3	rs738791_T rs2267029_A rs738792_C
MMP12	11	102733464 - 102745764	2	rs17368582_C rs11225442_A
MMP13	11	102813721 - 102826463	3	rs10502009_G rs3819089_A rs640198_A
MMP14	14	23305793 - 23316803	8	rs1042703_C rs762052_A rs8006914_T rs2236302_G rs1042704_A rs2236307_C rs743257_T rs17882342_D
MMP15	16	58059282 - 58080804	0	
MMP16	8	89049460 - 89339717	56	rs10504847_T rs2664369_G rs2664370_C rs17719609_C rs16877270_G rs1477908_C rs10103111_T rs2616493_C

				rs10098052_A rs2664346_C rs2616488_C rs6469206_G rs7826929_G rs2616506_C rs17663841_C rs977231_G rs2664352_C rs11782395_A rs1477916_T rs17664125_C rs13277637_T rs4961076_C rs9297422_C rs1382105_T rs1477917_G rs2664361_C rs16878818_T rs10099888_C rs7819728_A rs1996637_C rs1519938_G rs6981717_C rs2176771_C rs1519942_G rs12546847_C rs4961080_C rs13261974_A rs6469298_T rs17666351_G rs13261169_T rs1401861_A rs1879201_G rs17666490_T rs16880099_A rs7826477_T rs6994019_T rs16880416_T rs2222294_T rs7817382_G rs7834743_A rs7816934_C rs7000030_T

				rs3851539_G rs10504846_A rs10094702_C rs7835845_T
MMP17	12	132312941 - 132336316	11	rs3087864_G rs4964924_T rs4964927_T rs11246838_G rs6598163_A rs34515698_T rs7300198_C rs12099648_A rs9634312_A rs11613757_T rs11835665_A
MMP19	12	56229214 - 56236767	3	rs2242295_A rs2291267_A rs2291268_G
MMP20	11	102447566 - 102496063	17	rs2292730_A rs11225332_C rs1711399_C rs1711433_G rs10895322_G rs1711430_T rs1711427_C rs1784425_G rs1784424_A rs3781787_C rs3781788_T rs17098913_A rs10502005_A rs2280211_C rs11225344_A rs1962082_T rs2245803_A
MMP21	10	127455027 - 127464390	3	rs7922546_A rs10901424_T rs12775804_A

MMP23B	1	1567560 - 1570030	0	
MMP24	20	33814539 - 33864804	21	kgp4728036_A kgp4471741_A kgp6966600_G kgp481229_T kgp2046320_G kgp7289875_G kgp5576338_T kgp10149373_G kgp7633769_A kgp9807173_C kgp1472099_T rs12479765_A rs2425022_C rs6088776_C rs2247828_G rs2425024_C rs2254207_C rs11696548_T rs6060341_G rs7280_G rs2425032_C
MMP25	16	3096682 - 3110724	7	rs2247226_T rs10431961_T rs7199221_A rs1064875_T rs1064948_A rs11864930_A rs10438593_T
MMP26	11	5009424 - 5013659	1	rs2499958_A

MMP27	11	102562415 - 102576468	15	rs12099177_A rs2509010_T rs11607205_A rs1276289_A rs1276286_T rs2846723_C rs2846701_G rs2846703_G rs3809018_A rs17099425_G rs11225386_G rs11225388_G rs2846707_A rs1939015_G rs11225389_A
MMP28	17	34083269 – 34122640	1	rs3826404_G

bp: base pair; **SNP**: Single Nucleotide Polymorphism.

Data Considerations

We have taken a number of measures while preparing the data files for analysis. For example, Cox-MDR automatically rounds numbers to the nearest integer; hence, we rounded the continuous variables (age, overall survival (OS) time). Note that the rounded OS time was used to determine the 5-year survival status that was used in the GMDR 0.9 analysis. Additionally, at least Cox-MDR requires a complete dataset (i.e. no missing data). To address this data requirement, we only included the patients with complete clinical data (n=439) and SNPs/polymorphisms with zero missing genotype rate.

Code Extension and Runs

We added to the Cox-MDR code the ability to retrieve genotype and training balanced accuracy values for each Cox-MDR model, perform multiple runs at once, utilize random seed setting/loading (e.g. required for permutation testing purposes), and conduct permutation testing for any selected model. After these extensions, Cox-MDR code was tested to examine its features and to ensure that it worked correctly and produced the correct output.

As part of this study, the permutation testing Perl script included with GMDR 0.9 was extended. Specifically, lines 160 and 213 were edited to allow the setting of random seeds, adding the parameter "-seed=<long>" as specified by the GMDR-0.9 --help command-line argument output.

Cox-MDR and GMDR 0.9 programs were run in R[141,186] (R versions 3.5.0, 3.5.1, 3.6.2; Microsoft R Open version 3.5.1) and Java respectively. Cox-MDR and GMDR 0.9 analyses on large interaction datasets were performed in parallel to reduce computational time using the hardware and software systems at the Center for Health Informatics and Analytics (CHIA), Memorial University of Newfoundland. For Cox-MDR analyses in CHIA, we used manually set seeds (generated in R) in order to ensure parallel runs had different cross-validation partitioning despite starting at roughly the same time, as by default the random number generation of R sets a random seed based on system time.

Permutation Testing

For Cox-MDR permutation testing, an R function was written and run through R-Studio[187]. This function randomly shuffles specified columns of the input data and can be called multiple times to produce different shuffles of the dataset. This function was designed so that, similar to GMDR 0.9's permutation testing procedure**,** elements in shuffled columns still remained together in the same row, but the relationship between these elements and all remaining elements (i.e. SNP genotypes) was randomized. For the permutation testing procedure, the Cox-MDR or GMDR 0.9 method was applied to 1000 random shuffles of the data. The permutation testing procedure ran the Cox-MDR or GMDR 0.9 program using the same random seed as the run (i.e. the run that identified the top MDR model) being tested to ensure the same patients were in the same cross-validation folds between runs. After 1000 runs, the p-value was determined to be the

number of testing balanced accuracy (TBA) values for Cox-MDR, or average (among cross-validation folds) TBA values for GMDR 0.9, which were as high as or higher than the observed TBA value for the "top model" divided by the number of permutations (n=1000); if this value was <= 5%, then the top MDR model was deemed to be significant[99] (i.e. not likely to be detected by chance). The permutation testing procedure for GMDR 0.9 functions identically to that of Cox-MDR, except that GMDR 0.9 software uses the average TBA value among the 5 cross validation folds instead of the highest TBA value.

For larger datasets using GMDR 0.9 performing the permutation testing procedure on a desktop computer exceeded hardware resources. For these sets, permutation testing was performed on the CHIA computing cluster.

Part 2: ***Interactions among the SNPs of VEGF interaction network genes***

Table S3: Baseline characteristics for the 400 patients included in the VEGF interactome study

VEGF Project		
Variable	**N**	**%**
Age At Diagnosis	Median: 61; Range 21- 75 years	
Disease Stage		

I	77	19.25
II	165	41.25
III	126	31.50
IV	32	8.00
MSI Status		
Stable/MSI-low	350	87.5
MSI-high	50	12.5
Tumor Location		
Colon	264	66.00
Rectum	136	34.00
Baseline Radiation		
Adjuvant	100	25.00
Others	300	75.00
Baseline Chemotherapy		
Adjuvant	223	55.75
Others	177	44.25
DSS Time	Median: 14; Range: 0 -19 years	
#DSS Status		

Alive	309	77.25
Dead	91	22.75
##5-Year DSS Status		
Alive at 5 years	337	84.25
Dead at 5 years	63	15.75

MSI: Microsatellite instability; **DSS**: Disease Specific Survival. #Used in Cox-MDR, Kaplan-Meier and Cox regression analyses. ##Used in GMDR 0.9 and logistic regression analyses. Note that this table includes the 5 patients, who were removed from GMDR 0.9 analysis

Identification of Interaction Partners of the VEGF Family Genes

For the BioGRID[148–150] (; BioGRID | Database of protein, chemical, and genetic interactions) searches, species was set to "Homo sapiens". Interactions were downloaded in BioGRID TAB 2.0 format. Note that BioGRID uses these aliases for the following VEGF family proteins in their records: FLT1 for VEGFR1, KDR for VEGFR2, FLT4 for VEGFR3, and PGF for PIGF. Chemical interactions were filtered out of results of BioGRID searches before producing BioGRID TAB files and network diagrams (**Figure S1**). Non-human interactors were removed from the interactor datasets. A set of interactors for each VEGF family gene was produced by combining the columns "Official Symbol Interactor A" and "Official Symbol Interactor B" for the BIOGRID TAB files and removing duplicate gene symbols. The number of interactors of each VEGF gene are given in **Table S4A**.

In cases where multiple gene symbols had the same genomic location, we kept only the symbol which was found in our BioGRID search.

While using legacy Biomart section to obtain genomic locations for interactors, genes that were located on the X chromosome were excluded from further analysis (**Table S4A**). Additionally, entries with "PATCH" annotations were removed from analysis, as were entries with unusual annotations, opting for locations from the UCSC under the "Comprehensive Gene Annotation Set from GENCODE Version 19" heading. It was confirmed that the same gene in different interaction sets had the same genomic location.

Figure S1: Interaction networks for VEGFA, VEGFB, VEGFC, VEGFR1, VEGFR2, VEGFR3, and PIGF taken from Biogrid.

Figure S1 A: VEGFA interaction network.

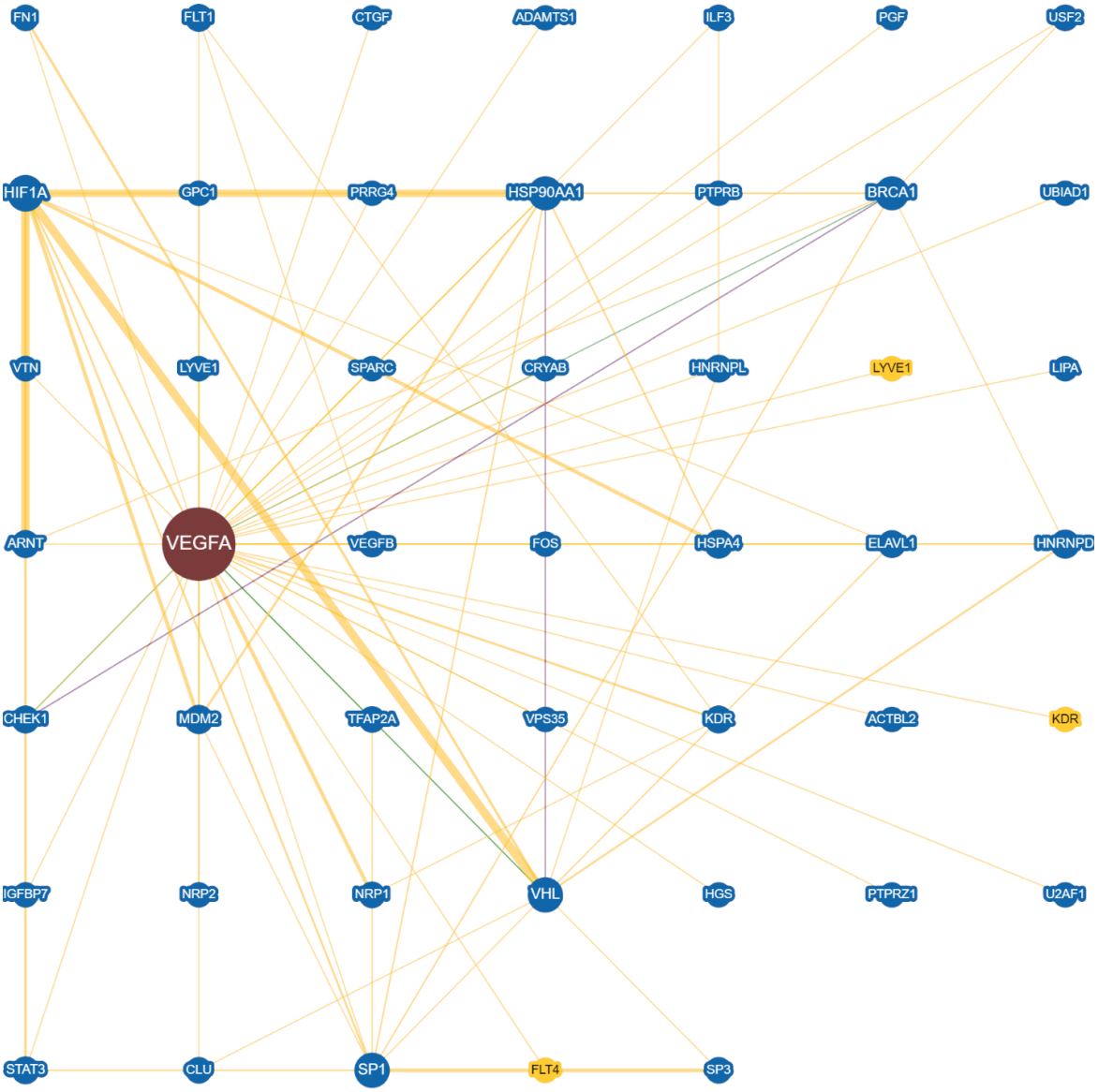


Figure S1 B: VEGFB interaction network.

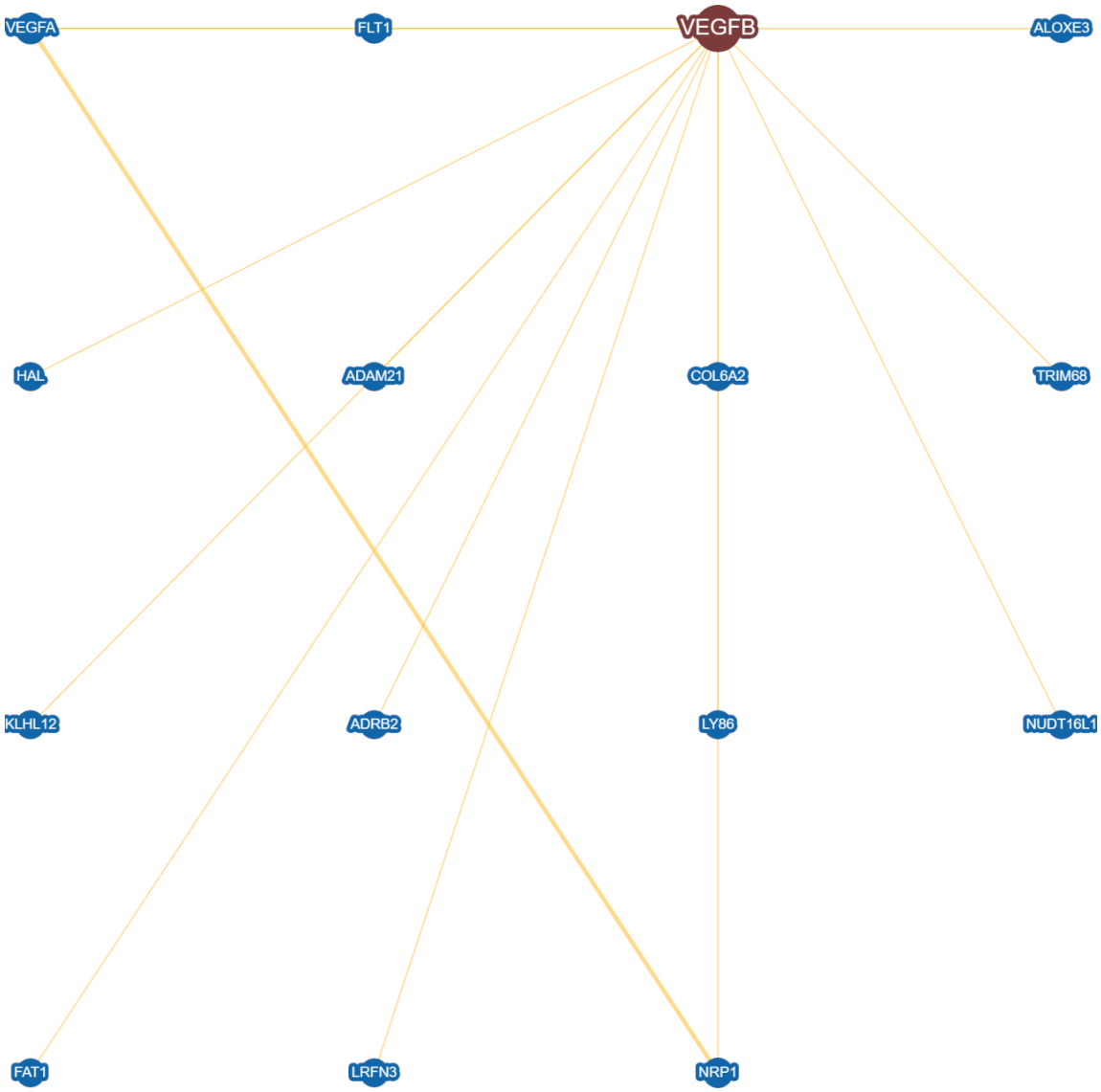


Figure S1 C: VEGFC interaction network.

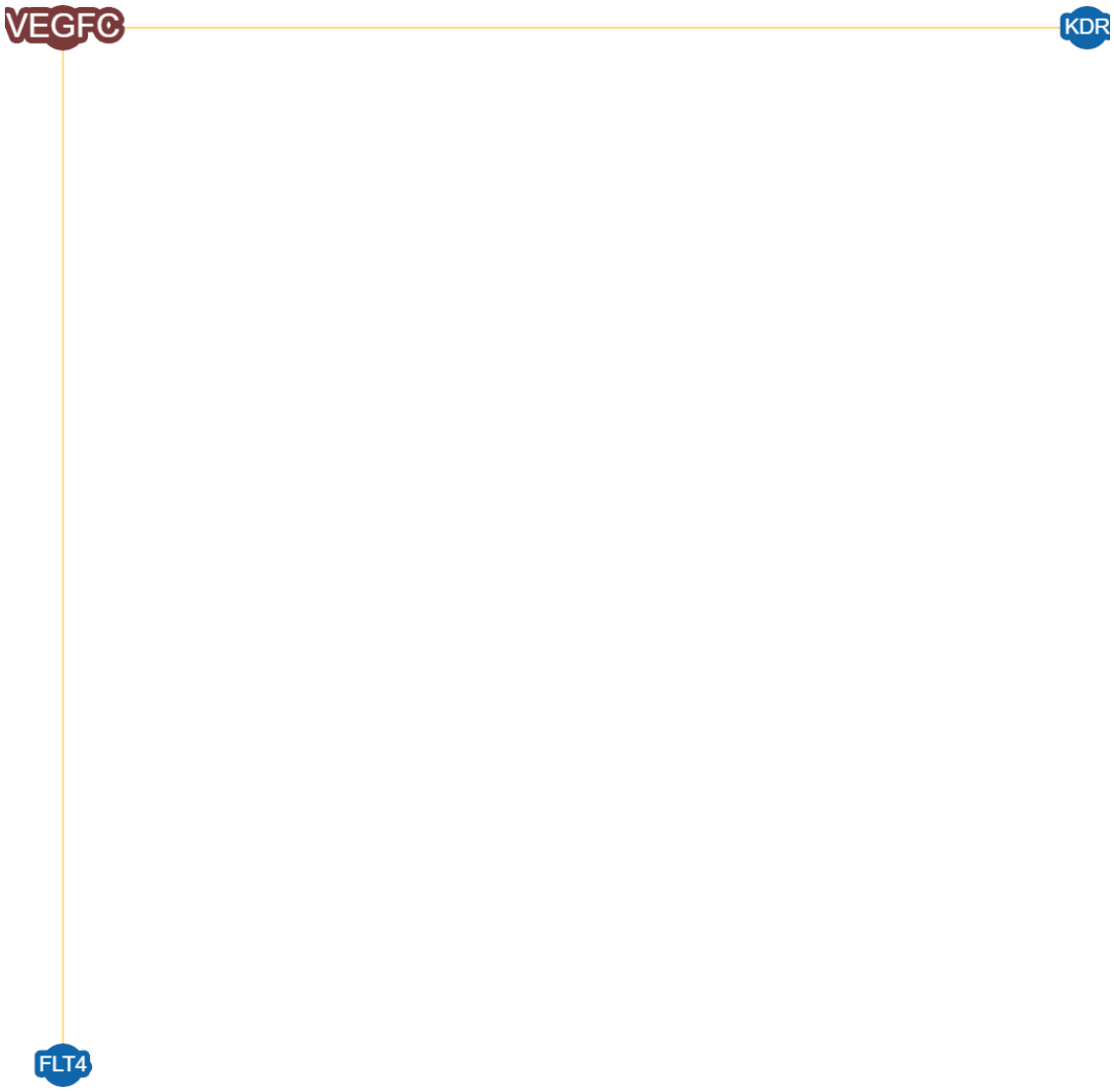


Figure S1 D: VEGFR1/FLT1 interaction network.

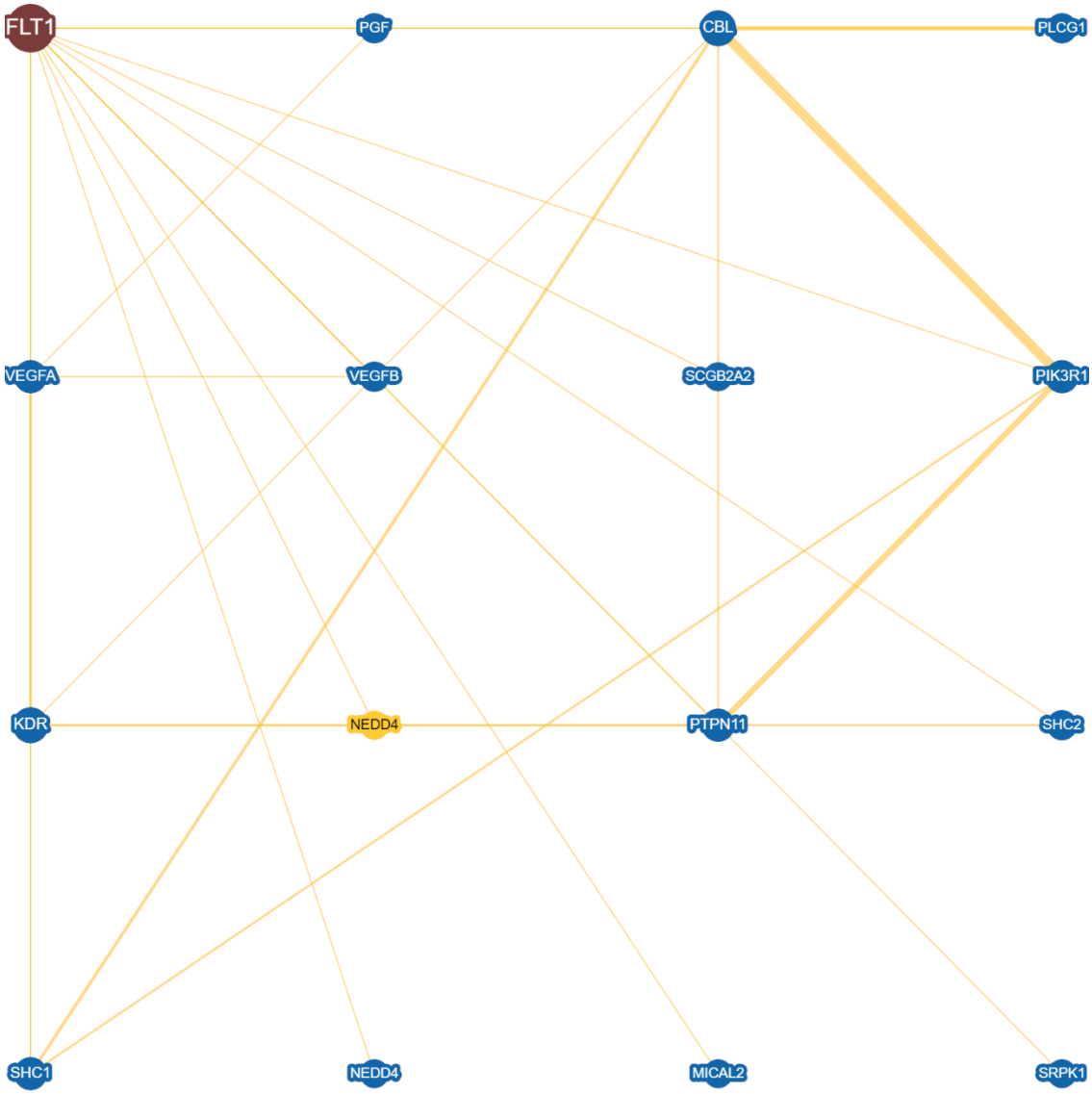


Figure S1 E: VEGFR2/KDR interaction network.

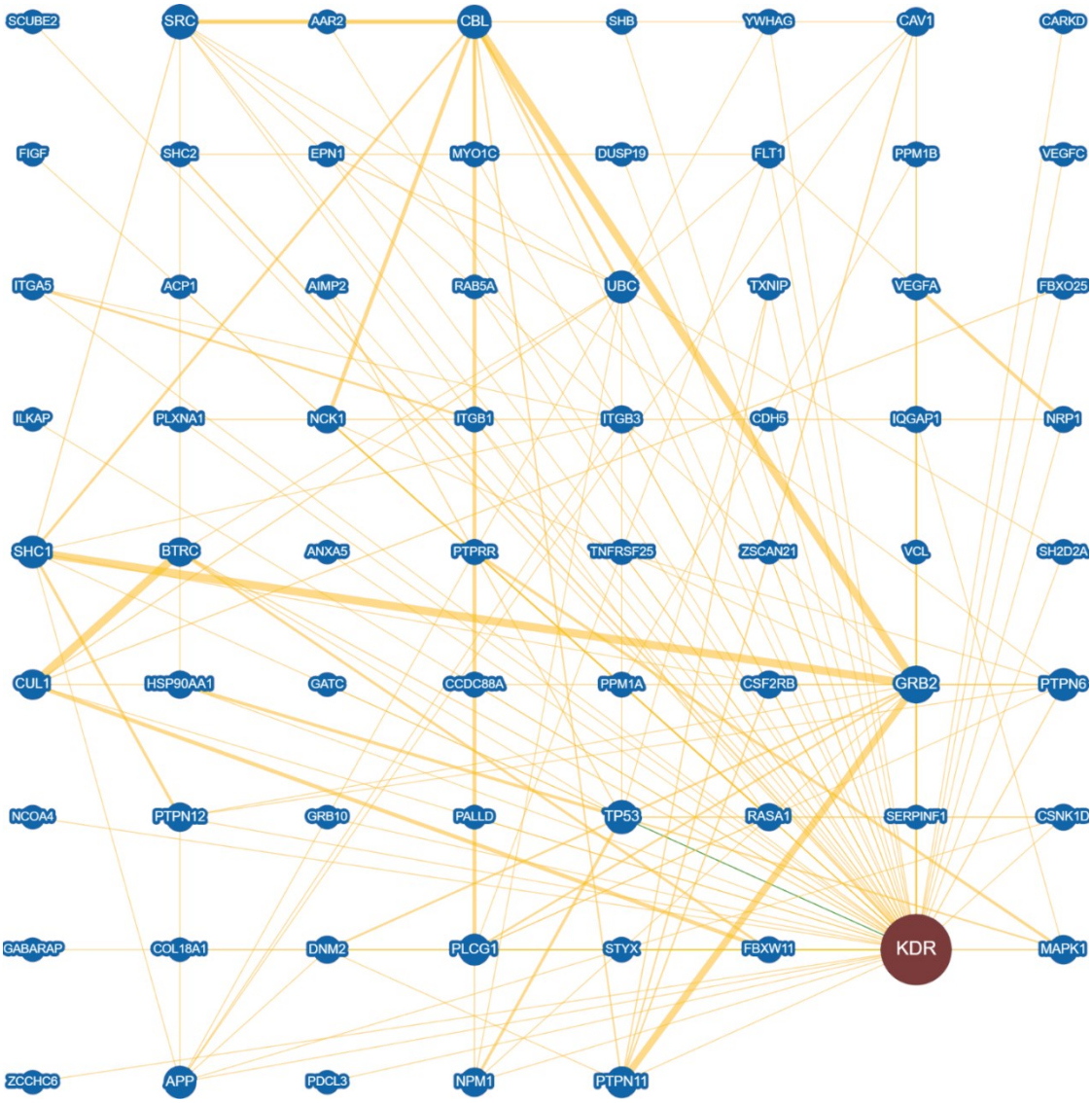


Figure S1 F: VEGFR3/FLT4 interaction network.

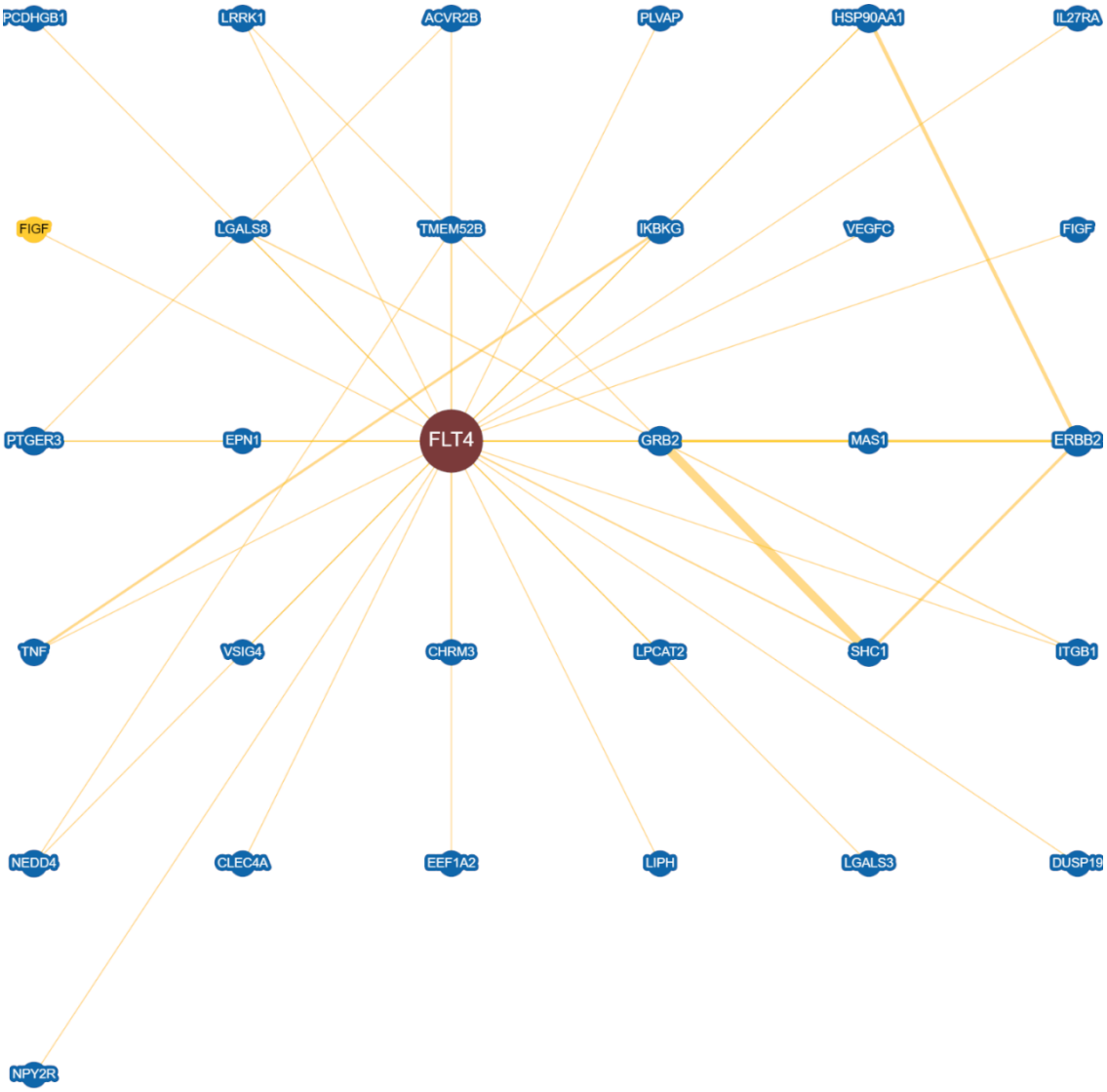
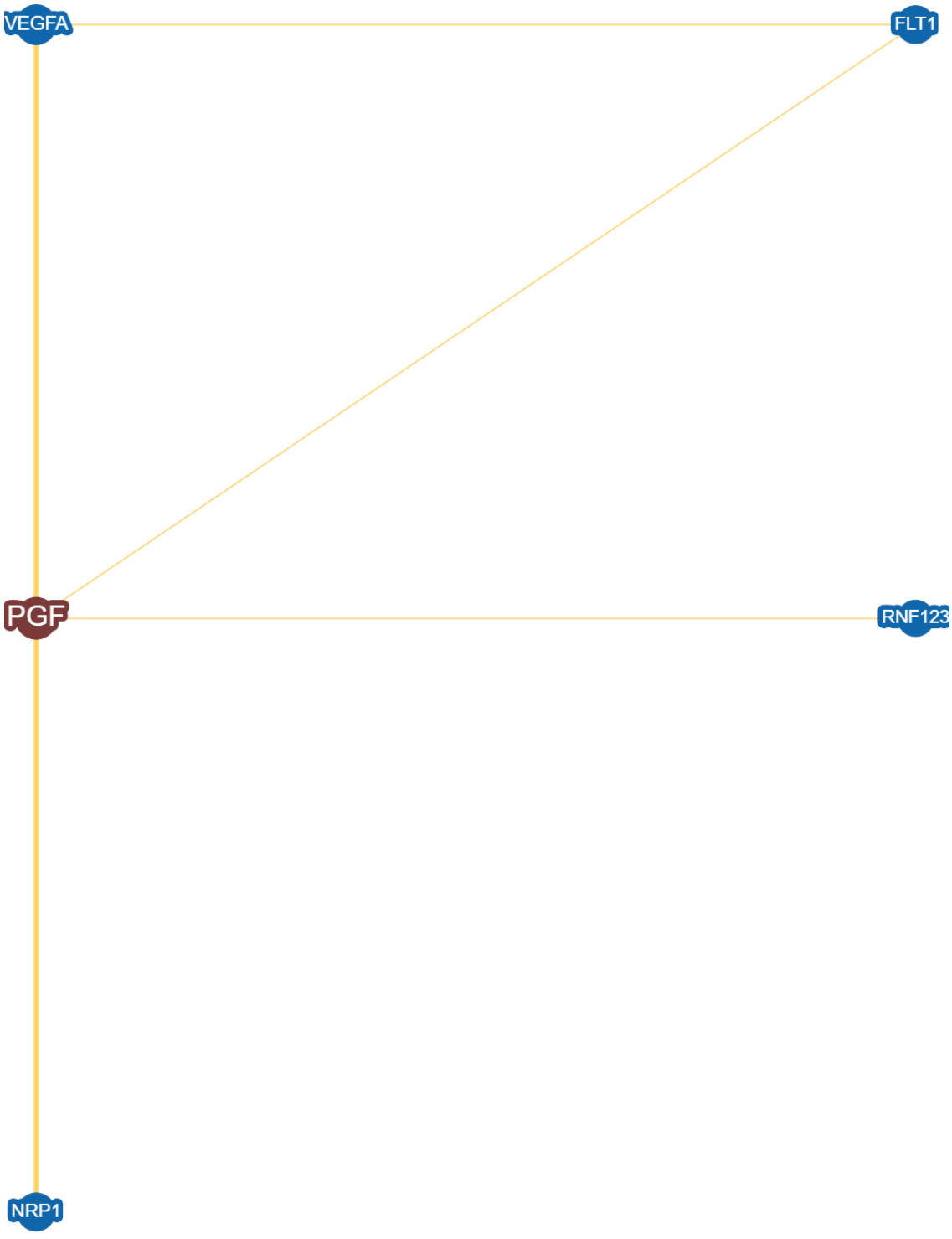


Figure S1 G: PIGF/PGF interaction network.



Network diagrams generated using BioGRID[149,150]. Line size represents the number of unique interactions in the BioGRID database for a given pair of interactors. Yellow lines represent physical interactions, green lines represent genetic interactions, and purple lines represent evidence of both physical and genetic interactions. Yellow nodes represent non-human genes which were subsequently removed from analysis. Aliases: FLT1, KDR, FLT4, and PGF refer to VEGFR1, VEGFR3, VEGFR3, and PIGF respectively.

Table S4 A: Number of VEGF interactors in each network.

VEGF family member protein	**Number of genes in the interaction network**	**Number of genes in the interaction network***
VEGFA	43	43
VEGFB	14	14
VEGFC	3	3
PIGF	5	5
VEGFR1	15	15
VEGFR2	69	68
VEGFR3	26	23
TOTAL	**175**	**171**

*After the genes on the X chromosome are excluded.

Table S4 B: Final list of genes included in study of VEGF family interactome networks

VEGF family member protein	Final list of genes included in study
VEGFA	ACTBL2 ADAMTS1 ARNT BRCA1 CHEK1 CLU CRYAB ELAVL1 FLT1(VEGFR1) FN1 FOS GPC1 HGS HIF1A HNRNPD HNRNPL HSP90AA1 HSPA4 IGFBP7 ILF3 KDR(VEGFR2) LYVE1 MDM2 NRP1 NRP2 PGF(PIGF) PRRG4 PTPRB PTPRZ1 SP1 SP3 SPARC STAT3 TFAP2A U2AF1 USF2 VEGFA VEGFB VHL VPS35 VTN
VEGFB	ADAM21 ADRB2 ALOXE3 COL6A2 FAT1 FLT1 HAL KLHL12 NRP1 TRIM68 VEGFA VEGFB
VEGFC	FLT4(VEGFR3) KDR(VEGFR2) VEGFC
VEGFR1	CBL FLT1(VEGFR1) KDR(VEGFR2) MICAL2 NEDD4 PGF(PIGF) PIK3R1 PLCG1 PTPN11 SCGB2A2 SHC2 SRPK1 VEGFA VEGFB
VEGFR2	AAR2 ACP1 AIMP2 ANXA5 APP BTRC CARKD CAV1 CBL CCDC88A CDH5 COL18A1 CSF2RB CSNK1D CUL1 DNM2 DUSP19 EPN1 FBXO25 FBXW11 FLT1(VEGFR1) GABARAP GATC GRB2 GRB10 HSP90AA1 ILKAP IQGAP1 ITGA5 ITGB1 ITGB3 KDR(VEGFR2) MAPK1 MYO1C NCK1 NCOA4 NPM1 NRP1 PALLD PDCL3 PLCG1 PLXNA1 PPM1A PPM1B PTPN6 PTPN11 PTPN12 PTPRR RAB5A RASA1 SCUBE2 SERPINF1 SH2D2A SHB SHC2 SRC STYX TNFRSF25 TP53 UBC VCL VEGFA VEGFC YWHAG ZCCHC6 ZSCAN21
VEGFR3	CHRM3 DUSP19 EEF1A2 EPN1 ERBB2 FLT4(VEGFR2) GRB2 HSP90AA1 ITGB1 LGALS3 LGALS8 LIPH LRRK1 MAS1 NEDD4 NPY2R PCDHGB1 PLVAP PTGER3 TMEM52B TNF VEGFC
PIGF	FLT1 NRP1 PGF(PIGF) RNF123 VEGFA

Parentheses indicate alternate gene symbols used in this document

Table S5: All genes and SNPs included in VEGF interactor study

Gene	Number of SNPs	SNPs
AAR2	3	rs2425193 rs2104007 rs2425202
ACP1	1	rs7584915
ACTBL2	1	rs13159014
ADAM21	3	rs11622815 rs2000352 rs4143920
ADAMTS1	4	rs9636786 rs13615 kgp10200667 rs370850
ADRB2	3	rs1042711 rs1042713 rs1042717
AIMP2	6	rs1860461 rs1860460 rs6979676 rs7803611 rs7781199 rs4560
ALOXE3	10	rs3809882 rs4792214 rs6503080 rs9894356 rs2289587 rs7215658 rs3027209 rs4414548 rs4792239 rs3027229
ANXA5	7	rs11098637 rs12511956 rs10518391 rs6534309 rs6857766 rs13145977 rs2306416
APP	45	rs214486 rs3787620 rs2829973 rs1876064 rs454017 rs1787438 rs17001492 rs1783016 rs214488 rs2829984 rs2234983 rs216779 rs367489 rs440666 rs2014146 rs216762 rs1701000 rs9305268 rs128647 rs2096488 rs8132200 rs12626960 rs7278838 rs2830008 rs7281216 rs768040 rs2070655 rs2830028 rs2830034 rs2830038 rs1041420 rs7283136 rs2830044 rs2070654 rs2830051 rs2830052 rs11702267 rs2830067 rs2830071 rs2830088 rs17588612 rs455465 rs458848 rs4817090 rs2830101
ARNT	5	rs10847 rs10305710 rs2228099 rs2134688 rs11204737
BRCA1	5	rs8176305 rs3737559 rs1799950 rs799923 rs799912
BTRC	9	rs7090670 rs10786634 rs7901883 rs4451650 rs9419913 rs9420839 rs17767748 rs4151060 rs11595968
CARKD	2	rs330550 rs179356
CAV1	10	rs926198 rs10256914 rs3807986 rs959173 rs3807989 rs3815412 rs1022436 rs9920 rs1049334 rs1049337
CBL	5	rs6589722 rs1893032 rs2511844 rs11217234 rs1052121
CCDC88A	9	rs2576692 rs1047948 rs1545121 rs4484068 rs3099084 rs10496042 rs11684805 rs6721972 rs6545492
CDH5	11	rs10852432 rs1077318 rs1076019 rs2344564 rs7499886 rs2344565 rs1130844 rs11640843 rs1073584 rs16956504 rs1972839
CHEK1	6	rs3731395 rs10893405 rs521102 rs2282535 rs11220181 rs11220182
CHRM3	94	rs4130463 rs10925877 rs12123857 rs6678395 rs12137225 rs10802767 rs6676664 rs6687984 rs17645304 rs12086449

		rs10925888 rs16838380 rs1782349 rs1782357 rs12041334 rs10925907 rs16838444 rs17646815 rs7525710 rs11804608 rs6670728 rs12135445 rs13376565 rs6429140 rs12119540 rs6429144 rs12120382 rs17594385 rs6685121 rs6688669 rs6690612 rs2355230 rs726168 rs12088787 rs12037424 rs6691263 rs10925941 rs12090480 rs10802789 rs1867266 rs1867265 rs6692711 rs12406493 rs4145784 rs2278642 rs1431718 rs12124903 rs10925971 rs685475 rs685550 rs685960 rs843030 rs6703930 rs7533134 rs17657156 rs2841037 rs481036 rs4266870 rs483411 rs693948 rs665159 rs1111249 rs12059546 rs558438 rs6690809 rs7543259 rs6429157 rs1578180 rs1934349 rs7536133 rs6698105 rs589962 rs685548 rs602117 rs1125489 rs1594513 rs10925994 rs682355 rs536477 rs2217533 rs10802812 rs16839034 rs16839045 rs10926008 rs16839051 rs10926009 rs7527677 rs10399860 rs12036109 rs7520974 rs6701181 rs6682184 rs4431831 rs4659554
CLU	4	rs9331947 rs7812347 rs7982 rs9331888
COL18A1	31	rs879330 rs8128168 rs2026886 rs4819099 rs4819101 rs2838916 rs2838917 rs2838920 rs2838923 rs8126757 rs11702782 rs7275991 rs9980531 rs4819115 rs2236451 rs11702494 rs2230688 rs2236459 rs2838942 rs2246749 rs11702425 kgp9623698 kgp383228 rs2236475 rs7279445 rs3753019 rs2236483 rs12483553 rs7278425 rs17004785 rs7867
COL6A2	10	rs9978018 rs2839108 rs17357592 rs2839112 rs2839113 rs7280485 rs2839116 rs3088026 rs1044598 rs2839117
CRYAB	3	rs4252588 rs11214040 rs11214043
CSF2RB	9	rs2075726 rs5756407 rs6000488 rs11089810 rs909486 rs1534882 rs11705394 rs131840 rs131842
CSNK1D	2	rs11653735 rs4789846
CUL1	13	rs243551 rs243538 rs243524 rs243523 rs11760399 rs243492 rs243482 rs243477 rs1014095 rs3823635 rs10271133 rs7779159 rs2007404
DNM2	10	rs12974306 rs4334414 rs714307 rs892086 rs12232826 rs4804524 rs7246673 rs2278444 rs2287029 rs12461992
DUSP19	7	rs16823976 rs3748880 rs12463411 rs11883456 rs2705730 rs17704934 rs2944346
EEF1A2	5	rs2274860 rs2750395 rs310619 rs8126435 rs910948
ELAVL1	8	rs2042920 rs12983784 rs4804244 rs759817 rs10401186 rs3786619 rs7251814 rs1204494
EPN1	7	rs8104242 rs3786642 rs10408454 rs10410404 rs6509955 rs7255531 rs2287831

ERBB2	6	kgp11187652 kgp8452497 kgp8195839 rs4252612 rs1801200 rs4252667
FAT1	31	rs3775309 rs1280092 rs2637777 rs1280103 rs28647489 rs2249916 rs2249917 rs3733406 rs11931107 rs189031 rs7663350 rs328432 rs328431 rs1388297 rs7672047 rs162062 rs167853 rs2130910 rs10155467 rs2130909 rs2375889 rs162182 rs4862723 rs327080 rs455600 rs907986 rs455219 rs13123522 rs3733414 rs1491248 rs327102
FBXO25	15	rs13279681 rs17812876 rs17665364 rs17064974 rs13340594 rs1530662 rs3735925 rs10092971 rs2278765 rs13253643 rs10088894 rs12546599 rs10503146 rs10109251 rs17665621
FBXW11	3	rs9313563 rs9313564 rs9313565
FLT1	36	rs9554314 rs12429309 rs9513070 rs12877323 rs3794397 rs3794399 rs2296188 rs2296189 rs7987291 rs7987649 rs3794400 rs2387632 rs3936415 rs17086609 rs1853581 rs7989623 rs7995976 rs9551462 rs3751395 rs17086617 rs17537350 rs7332329 rs9508021 rs9513099 rs11149523 rs9508034 rs9513112 rs9554330 rs3794405 rs9513113 rs10507386 rs585421 rs622227 rs655024 rs679791 rs598945
FLT4	16	rs307822 rs2279622 rs11739750 rs2242217 rs400330 rs3797104 rs307823 rs3797102 rs3736061 kgp53910 rs2290983 rs10085025 rs4700745 rs10072977 rs11748431 rs307814
FN1	18	rs1263 rs11651 kgp9543736 rs2289200 rs6707530 rs7608342 rs13652 rs1250201 rs7588661 rs11883812 rs1561302 rs7596677 rs17516906 rs724617 rs1437799 rs16854041 rs7609476 rs1250246
FOS	2	rs7101 rs1063169
GABARAP	2	rs11656323 rs222843
GATC	4	rs17431446 rs2235217 rs7957424 rs3847971
GPC1	13	rs7577243 rs13424854 rs7589322 rs3828334 rs3828336 rs2292832 rs881029 rs12695020 rs2228327 rs1126920 rs13013933 rs3792215 rs1042823
GRB10	32	rs4245556 rs4947406 rs4947709 rs2715129 rs11770199 rs17544225 rs2250152 rs2299150 rs980716 rs6948959 rs2715117 rs17544971 rs2237444 rs6593077 rs2237447 rs17133917 rs2237456 rs1800504 rs2237477 rs2237482 rs10248619 rs2282930 rs2299155 rs17152102 rs2108349 rs6968827 rs1024532 rs6979369 rs7805310 rs6976572 rs7791286 rs6593185
GRB2	3	rs16967789 rs959260 rs4789182

HAL	14	rs1059845 rs2230885 rs11108358 rs7297245 rs10492228 rs3213737 rs12319274 rs12307364 rs10745747 rs17676826 rs10859997 rs10492227 rs2302629 rs2302628
HGS	3	kgp785391 rs6565620 rs34384005
HIF1A	5	rs2301106 rs10129270 rs4899056 rs12434438 rs2057482
HNRNPD	3	rs2288338 rs1820577 rs1365872
HNRNPL	3	rs10403012 rs2278012 rs862456
HSP90AA1	5	rs7155973 rs3736807 rs11621560 rs10873531 rs1190603
HSPA4	3	rs13161158 rs11749966 rs14355
IGFBP7	35	rs1277308 rs11573128 rs2271808 rs1277311 rs11133472 rs1718885 rs7687211 rs6852762 rs3821996 rs6554404 rs881382 rs1713973 rs1401189 rs1713963 rs11573086 rs1713959 rs7656865 rs1277293 rs7356193 rs17761305 rs1718856 rs4865174 rs10516163 rs11934877 rs1714014 rs1718848 rs1714011 rs1718845 rs11936912 rs10019698 rs1718858 rs1718861 rs6851308 rs4865181 rs10004910
ILF3	2	rs2569507 rs13465
ILKAP	6	rs2278737 rs2880132 rs2880131 rs6431588 rs2305171 rs3795903
IQGAP1	7	rs17176602 rs6496674 rs12912995 rs16974212 rs11853271 rs9944285 rs3539
ITGA5	3	rs7306692 rs1270919 kgp6380544
ITGB1	16	rs2153875 rs2488320 rs2230396 rs3780873 rs10763902 rs10827163 rs10827164 rs1009002 rs11009157 rs1187078 rs2457705 rs1187095 rs2475193 rs10827167 rs1187086 rs11591508
ITGB3	12	rs10514919 rs7209700 rs11868894 rs2292867 rs8073229 rs5918 rs2292699 rs12603582 rs3809863 rs7225700 rs12948299 rs11867160
KDR	14	rs12642307 rs2125489 rs1531289 rs17709898 rs6838752 rs6828477 rs11732292 rs17085326 rs2034965 rs17711073 rs2305948 rs7692791 rs6837735 rs12502008
KLHL12	3	rs12089566 rs4950887 rs2275734
LGALS3	2	rs7160523 kgp43854
LGALS8	15	rs17753447 rs1266381 rs10802546 rs4659682 rs10925157 rs1266384 rs12041958 rs2799426 rs10925158 rs3754245 rs2472126 rs11807205 kgp6759139 rs2298096 rs2298098
LIPH	3	rs6788865 rs9790230 rs4626118
LRRK1	38	rs12148466 rs11630691 rs11858394 rs4075387 rs7170683 rs12441903 rs4965738 rs4965741 rs721906 rs8038607 rs12915954 rs7176253 rs2412000 rs12914811 rs12439038 rs966293 rs11857262 rs6598411 rs1993375 rs12595297

		rs878274 rs6598412 rs1078513 rs2034809 rs963333 rs930847 rs11633278 rs11247253 rs4427776 rs12594881 rs12592409 rs4965778 rs4965780 rs11857803 rs17161155 rs17744500 rs2925202 rs1048327
LYVE1	12	rs17403620 rs17318858 rs17318955 rs17403977 rs16907989 rs7111477 rs11042883 rs11042889 rs11042892 rs16927077 rs10840444 rs1017275
MAPK1	8	rs2276008 rs9340 rs17821423 rs2298432 rs2006893 rs9607272 rs17759796 rs8141815
MAS1	1	rs220721
MDM2	3	rs937283 rs2279744 rs1470383
MICAL2	94	rs11022172 rs7111481 rs7130896 rs12803936 rs2015963 rs3763820 rs2307072 rs4756772 rs10765923 rs12577615 rs7932017 rs7942252 rs12795108 rs12790969 rs10741566 rs10765924 rs12577704 rs11022188 rs7940840 rs977244 rs2171150 rs9971381 rs901284 rs11022193 rs988189 rs4757237 rs4756775 rs11022209 rs10831742 rs7102041 rs9804570 rs1564946 rs1564947 rs923167 rs7131034 rs901302 rs2010463 rs11022214 rs10831744 rs4471395 rs7950540 rs7121956 rs7130607 rs7949360 rs6485561 rs11022242 rs10831758 rs1032151 rs17477991 rs2013262 rs3763822 rs10430830 rs12283453 rs871703 rs2279390 rs12787479 rs954428 rs11022250 rs2012580 rs2306729 rs11827638 rs12294182 rs7101833 rs10831769 rs6485587 rs7103040 rs2706643 rs2641941 rs2706645 rs1609930 rs11022257 rs2010576 rs2246778 rs3794083 rs2706637 rs4757276 rs11022262 rs2641938 rs11604904 rs2279613 rs2270511 rs12574429 rs2706627 rs1973386 rs7946327 rs1493959 rs1826608 rs11022264 rs17480838 rs7116182 rs2270513 rs3794075 rs2279616 rs8808
MYO1C	7	rs2302459 rs2302458 rs2286870 rs2302456 rs2286873 rs2286876 rs7218128
NCK1	4	rs9845460 rs1347209 rs3772388 rs1048145
NCOA4	5	rs10761581 rs10740051 rs17720205 rs41306524 rs11548236
NEDD4	15	rs11550869 rs2899593 rs17238468 rs12898589 rs8031043 rs12232351 rs2414448 rs8027843 rs10518827 rs12593446 rs12591210 rs7174459 rs12592220 rs9920283 rs16976661
NPM1	1	rs11134696
NPY2R	3	rs17376826 rs1574175 rs1047214
NRP1	53	rs1044268 rs1044210 rs2506141 rs2506143 rs2506145 rs2228638 rs2383984 rs734186 rs2474723 rs11009281 rs2474712 rs2254826 rs2269096 rs1331317 rs11009311 rs927099 rs11009313 rs12765284 rs2269091 rs12762312

		rs17413155 rs17413169 rs10490939 rs1888688 rs11009323 rs2383987 rs1319013 rs11593943 rs3780869 rs10490938 rs16934292 rs11598845 rs2073320 rs4934584 rs17296436 rs17296443 rs10827227 rs10827228 rs869636 rs7079372 rs2776928 rs1331326 rs6481844 rs7910405 rs2776930 rs2776932 rs2065364 rs2804492 rs2804493 rs2776937 rs4934597 rs1360457 rs2804498
NRP2	37	rs10090 rs698909 rs849530 rs950219 rs849556 rs3771051 rs3771048 rs3771044 rs849542 rs3771038 rs861079 rs3771033 rs849523 rs849582 rs849575 rs849570 rs3771021 rs849565 rs849563 rs1996412 rs12472412 rs13026243 rs849560 rs2241155 rs3771016 rs872943 rs3771004 rs16837637 rs3771003 rs16837641 rs2241153 rs3732088 rs2160328 rs3771000 rs3770996 rs3755232 rs1990708
PALLD	104	rs11132268 rs2712135 rs2712149 rs13150330 rs2002727 rs4692943 rs10517996 rs1962363 rs10517999 rs6552861 rs10518001 rs6857497 rs7673220 rs17054290 rs11735275 rs11132283 rs9312333 rs13145788 rs1986369 rs6836618 rs6857016 rs11132322 rs17541413 rs10518011 rs10022002 rs10004025 rs1962022 rs7668720 rs17650886 rs6815330 rs17650892 rs2874112 rs17707379 rs2319909 rs4144994 rs4371580 rs4389538 rs12647503 rs2710850 rs17707568 rs2723687 rs2710851 rs2723688 rs3109799 rs2723696 rs2723698 rs12643131 rs2710828 rs2723704 rs17614077 rs10010321 rs12642267 rs17054449 rs4314247 rs12649186 rs2723705 rs13137200 rs6832582 rs17054460 rs4260495 rs4280700 rs12649675 rs4635780 rs9884230 rs4599370 rs7697688 rs11132434 rs17542430 rs1500800 rs6852874 rs12510359 rs17542654 rs17708307 rs4692948 rs7679564 rs2247733 rs999958 rs7688994 rs17614733 rs11733873 rs1875297 rs1875296 rs7681510 rs13129779 rs6854137 rs1566499 rs6854037 rs2133911 rs867901 rs973990 rs12643033 rs1318822 rs4692552 rs7688533 rs4692553 rs13114906 rs6852229 rs2062589 rs7682426 rs867632 rs12643097 rs2047633 rs6819031 rs1500795
PCDHGB1	22	rs17097231 rs13171859 rs4151698 rs11575956 rs3806832 rs4151699 rs6867460 rs3749770 rs4912750 rs11575963 rs11958830 rs1423148 rs3805695 rs11748256 rs13361997 rs1002519 rs11952292 rs2237079 rs11744379 rs4912762 rs17286954 rs970069
PDCL3	4	rs6747613 rs2946589 rs12469806 rs2970997
PGF	2	rs8185 rs12411

PIK3R1	24	rs171648 rs7701498 rs831227 rs706713 rs13173003 rs7709243 rs12652661 rs173704 rs173702 rs4122269 rs1823023 rs173703 rs6893676 rs34303 rs863818 rs34309 rs2302975 rs3730082 rs6876003 rs3815701 rs34306 rs1550805 rs831125 rs3730089
PLCG1	4	rs2866370 rs753381 rs6072299 rs4297946
PLVAP	4	rs4808078 rs7252581 rs16981755 rs10417806
PLXNA1	7	rs6764158 rs732737 rs747967 rs9289290 rs4679325 rs9851451 rs3749395
PPM1A	3	rs7155841 rs10142834 rs12434739
PPM1B	4	rs1453863 rs17039151 rs4952703 rs2053456
PRRG4	7	kgp11715177 rs33962176 rs11605633 rs7944652 rs11032017 kgp8085505 rs7933966
PTGER3	49	rs959 rs1327460 rs6656853 rs6672081 rs7530738 rs7533733 rs6685546 rs6685646 rs17481440 rs1536537 rs1536261 rs1576055 rs4649932 rs35702222 rs1409166 rs1409165 rs1327464 rs1409162 rs4420040 rs7530658 rs2182325 rs11209714 rs875727 rs17541722 rs7539384 rs17542063 rs6424410 rs7538034 rs6670616 rs12067140 rs510414 rs475468 rs1409984 rs1071020 rs571705 rs977214 rs2072947 rs479934 rs2206343 rs2268055 rs2300168 rs5693 rs5691 rs1022528 rs8179390 rs2300179 rs10889906 rs2050065 rs2817867
PTPN11	4	rs11066301 rs17822304 rs12423190 rs11066323
PTPN12	8	rs9886084 rs10808113 rs2286894 rs1024723 rs7776973 rs17381884 rs17467232 kgp4610958
PTPN6	5	rs2301262 rs10774452 rs2110071 rs2071079 rs759052
PTPRB	42	rs431716 rs630608 rs17226367 rs2278346 rs17226374 rs2567142 rs919594 rs2567140 rs11178281 rs3782377 rs2567137 rs2584026 rs2567133 rs2116209 rs12314266 rs4761222 rs2303963 rs2717440 rs7954837 rs991833 rs2116211 rs2034011 rs2304821 rs2717418 rs2584011 rs2165627 rs11178317 rs2465811 rs11178321 rs2717430 rs2583999 rs751363 rs2439732 rs2465810 rs10506598 rs7298147 rs2717425 rs11178333 rs17814416 rs17108441 rs1442205 rs2717417
PTPRR	47	rs2717445 rs10879175 rs7298378 rs7314925 rs11178364 rs12580224 rs2089975 rs1398602 rs6581958 rs1156461 rs11178376 rs12813125 rs972769 rs1398599 rs11178388 rs3803036 rs7974346 rs4760933 rs6581964 rs1513098 rs7968934 rs7297717 rs7306190 rs4760744 rs4760810 rs10879198 rs1022242 rs17814482 rs2048607 rs7956670 rs2203232 rs17108861 rs10784870 rs12229663 rs4294640

		rs12305560 rs12297391 rs10879213 rs17108998 rs3923909 rs10879214 rs11178478 rs4760847 rs3924187 rs6581971 rs7965899 rs4595639
PTPRZ1	22	rs740965 rs1007784 rs12669706 rs1019221 rs960930 rs6466808 rs6970897 rs2690271 rs1196510 rs13246377 rs1209633 rs3817483 rs1196505 rs2693657 rs1196473 kgp11436861 rs1147504 rs1147498 rs1147492 rs1147491 rs1147489 rs1147487
RAB5A	9	rs11128928 rs4858660 rs17181547 rs4241539 rs2127956 rs4398451 rs9835991 rs13085694 rs8682
RASA1	5	rs6452750 rs35148638 rs10045850 rs2923742 rs10057748
RNF123	4	rs11130216 rs1491985 kgp9864706 rs11130218
SCGB2A2	1	rs17709552
SCUBE2	31	rs1136966 rs1367 rs2056902 rs3751057 rs7109896 rs3794149 rs1883100 rs10743098 rs1883099 rs10840164 rs6486112 rs3751055 rs10769988 rs7106593 rs3751051 rs3763904 rs7112378 rs7130913 rs6486125 rs2003906 rs4910431 rs7107892 rs11606516 rs7929797 rs2647528 rs1121629 rs11042182 rs3898554 rs4910443 rs10769990 rs10769992
SERPINF1	5	rs11658342 rs1136287 rs12603825 rs8074840 rs6828
SH2D2A	2	rs926103 rs2150906
SHB	12	rs776023 rs776022 rs776015 rs735740 rs3827519 rs3802414 rs10973635 rs12345885 rs7047051 rs7856790 rs943936 rs4878743
SHC2	11	rs8902 rs1046822 rs16990450 rs10426188 rs12981152 rs10408164 rs8112380 kgp471423 rs10409912 rs4919871 rs740871
SP1	2	rs3741651 rs17695156
SP3	3	rs6711060 rs4508563 rs10190140
SPARC	8	rs707156 rs3210714 rs729853 rs725937 rs2881558 rs17718347 rs11745387 rs17112187
SRC	13	rs7275012 rs16986606 rs6017996 rs6018027 rs6063022 rs6018088 rs6090575 rs12329503 rs6090585 rs754625 rs6018257 rs1570209 rs17785475
SRPK1	3	rs17704843 rs3761981 rs11968721
STAT3	6	rs1053005 rs1053004 rs8069645 rs6503695 rs744166 rs4796791
STYX	3	rs10483617 rs11625099 rs10873061
TFAP2A	6	rs537112 rs533558 rs303050 rs3798696 rs1675414 rs303055
TMEM52B	8	kgp11971666 rs7315498 rs7305138 rs10505752 rs12313003 rs17808107 rs4764306 rs4764308

TNF	1	rs3093662
TNFRSF25	1	rs11800462
TP53	7	rs8073498 rs12949853 rs1614984 rs1625895 rs1042522 rs8079544 rs11652704
TRIM68	2	rs3750992 rs931811
U2AF1	3	rs3788054 rs4920039 rs1789956
UBC	2	rs41276688 rs13624
USF2	3	rs2515622 kgp22836814 rs10405246
VCL	10	rs12250729 rs4746166 rs10458640 rs10458657 rs11000851 rs11000864 rs11000869 rs767809 rs2279648 rs3793921
VEGFA	7	rs25648 rs833068 rs833070 rs3024994 rs3025010 rs3025039 rs3025053
VEGFB	2	rs11603042 rs4930152
VEGFC	8	rs2877961 rs17697359 rs1485762 rs1485766 rs11947611 rs3775195 rs2171083 rs4557213
VHL	1	rs1642742
VPS35	1	rs700582
VTN	3	rs2277667 kgp4183944 rs2071379
YWHAG	2	rs2908191 rs917424
ZCCHC6	4	rs7035034 rs4587414 rs10115526 rs700759
ZSCAN21	2	rs11558475 rs12705070

Genes and SNPs used in the Part 2 (VEGF interaction network) analyses. Genes are listed in alphabetical order.

LD pruning: PLINK was used for genotype extraction, followed by LD-based pruning (using the PLINK the command *indep-pairwise* with a window of 50 SNPs, a step size of 5 SNPs[188,189] and a threshold of 0.8 (LD > 0.8 removed) was used.

RESULTS

Table S6: Results of the 1-way Cox-MDR runs (n=20) examining the MMP gene SNPs (n=201).

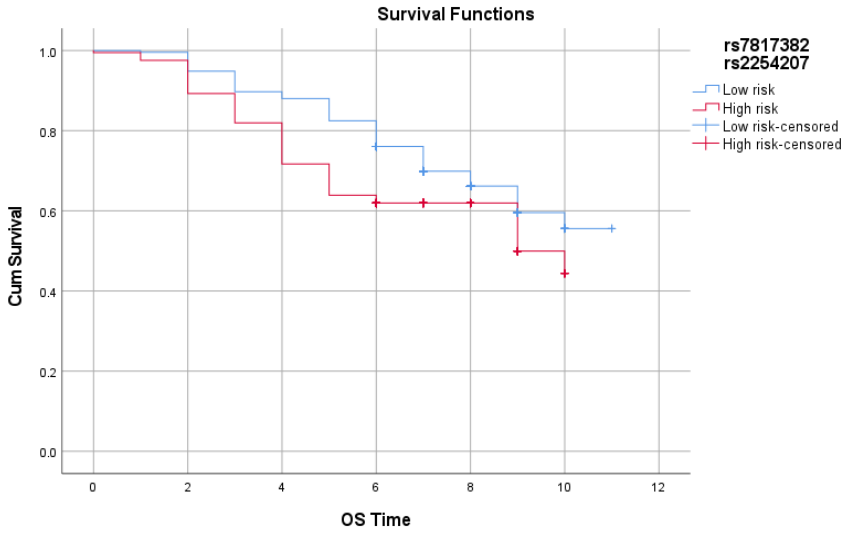
Run #	CVC	Testing Balance Accuracy	Best SNP	Genotype Risk Categorization (High Risk Shown)
10	2	0.512963	rs11225388	0:230
12	4	0.590351	rs11225388	0:230
17	4	0.595105	rs11225388	0:230
6	3	0.595118	rs11225388	0:230
5	5	0.596491	rs11225388	0:230
13	3	0.597114	rs11225388	0:230
4	4	0.598295	rs11225388	0:230
3	3	0.610472	rs11225388	0:230
1	3	0.613158	rs11225388	0:230
7	4	0.616848	rs11225388	0:230
16	4	0.617092	rs11225388	0:230
19	4	0.62069	rs11225388	0:230
11	4	0.62193	rs11225388	0:230
20	5	0.630688	rs11225388	0:230
8	4	0.642281	rs11225388	0:230

2	4	0.643003	rs11225388	0:230
9	4	0.64537	rs11225388	0:230
15	4	0.648148	rs11225388	0:230
18	4	0.649482	rs11225388	0:230
14	**5**	**0.656477**	**rs11225388**	**0:230**

CVC: Cross-Validation Consistency. Using our selection procedure, rs11225388, with AA genotype being the high risk genotype, AG and GG genotypes being the low risk genotypes, was found to be the most frequent (and top) MDR model. Genotype risk classification format: SNP Genotype: Number of patients in genotype risk category. Genotypes are presented with additive coding (0=major allele homozygous genotype; 1=heterozygous genotype, 2=minor allele homozygous). The top model is bolded. Data in this table is sorted by TBA and genotype risk categorization.

Figure S2. Kaplan Meier curves of the models identified by GMDR 0.9 in MMP SNP interaction analysis

A. 2-way model:

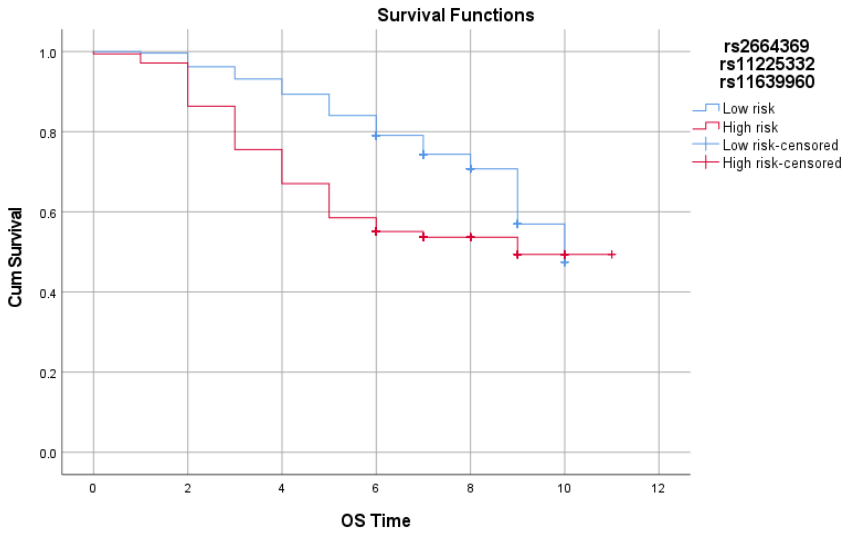


Log-rank p= 0.0152154116604927

Red: (AA,CA),(AA,CC),(GA,AA),(GA,CC),(GG,CA)

Blue: All other genotype combinations

B. 3-way model:



Log-rank p = 0.0000209959191927817

Red: (0TT,0TT,2GG), (0TT,1CT,1GA), (0TT,1CT,2GG), (0TT,2CC,1GA), (1GT,0TT,0AA), (1GT,0TT,1GA), (1GT,1CT,2GG), (1GT,2CC,2GG), (2GG,0TT,0AA), (2GG,1CT,2GG), (2GG,2CC,0AA), (2GG,2CC,2GG)

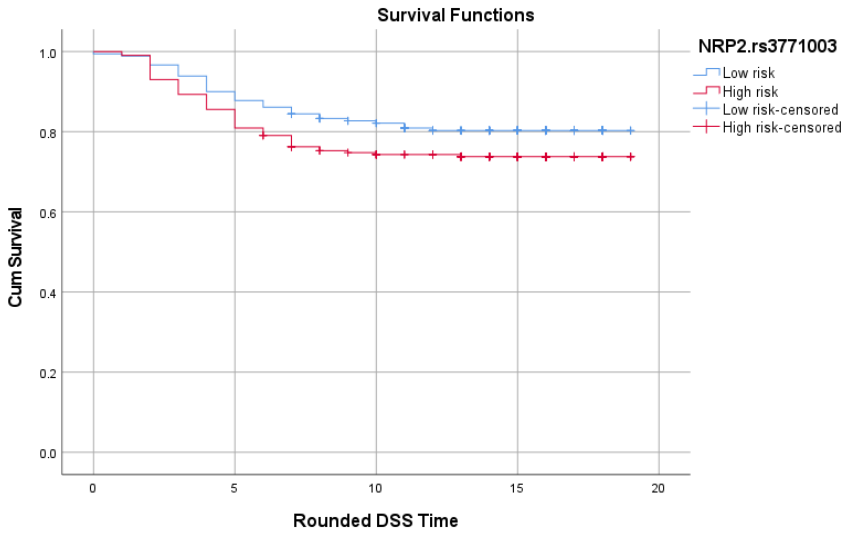
Blue: All other genotype combinations

Figure S3. Kaplan-Meier curves for models identified in the VEGF interaction network analysis by GMDR 0.9.

Red: high risk genotypes, blue: low risk genotypes

VEGFA

1-way model, iteration 1:

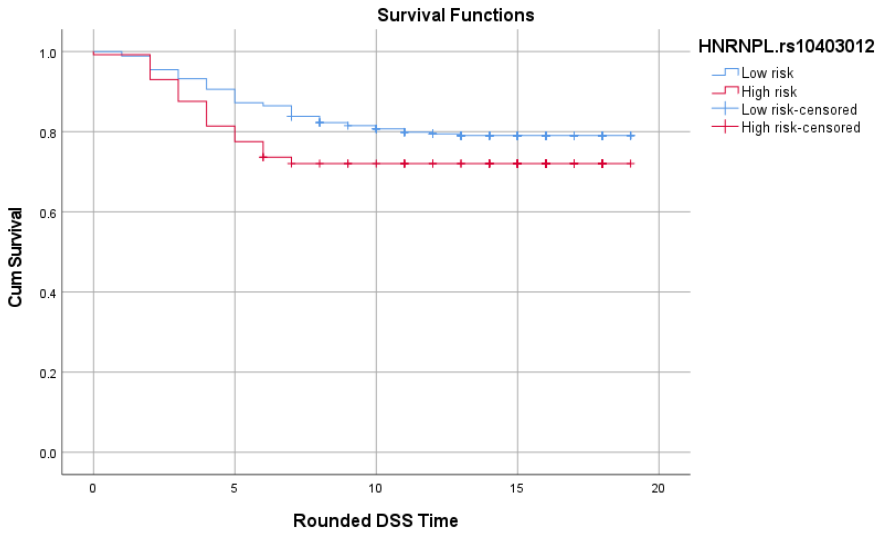


Log-rank p = 0.1064349793977

Red: GG and TT

Blue: TG

1-way model, iteration 2:

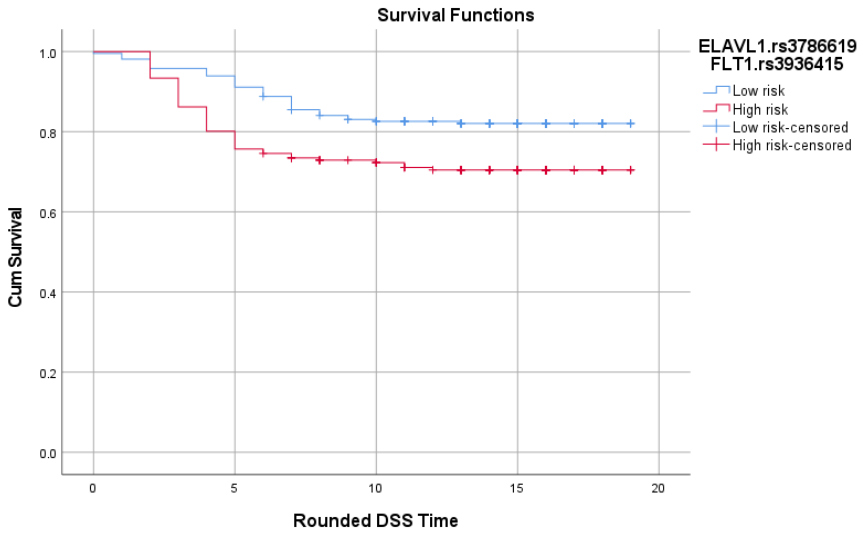


Log-rank p = 0.0749006184227615

Red: AA

Blue: GA, GG

2-way model:

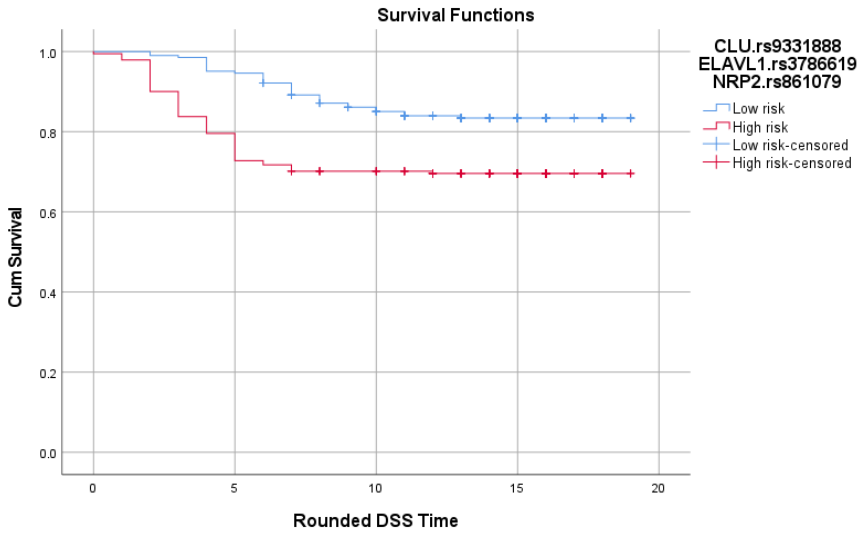


Log-rank p = 0.00393290069799821

Red: (GG,AA), (AG,GG), (AA,GG), (AA,AG)

Blue: All other genotype combinations

3-way model:



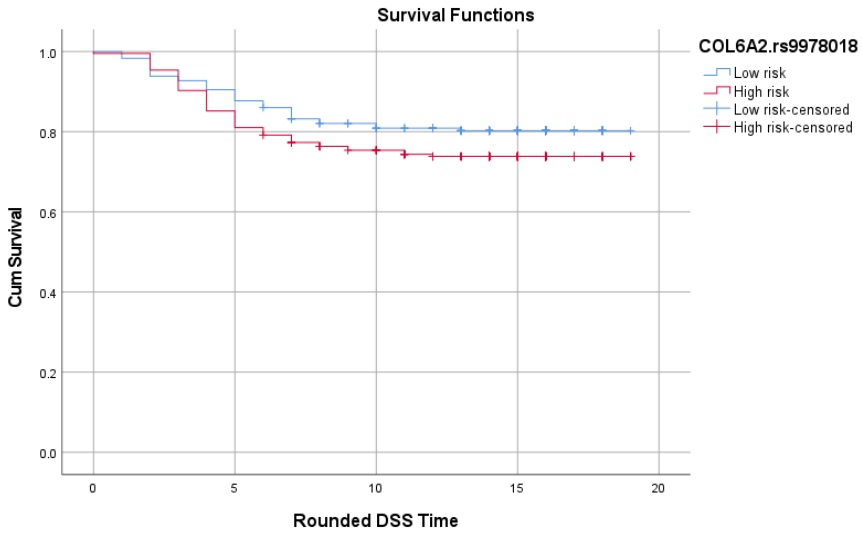
Log-rank p = 0.000238697896697943

Red: (GG,GG,TC), (GG,GG,TT), (GG,AG,TC), (GG,AA,CC), (GG,AA,TT), (CG,AG,CC), (CG,AG,TT), (CG,AA,CC), (CC,GG,TC), (CC,GG,TT), (CC,AG,TT), (CC,AA,TT)

Blue: All other genotype combinations

VEGFB

1-way model:

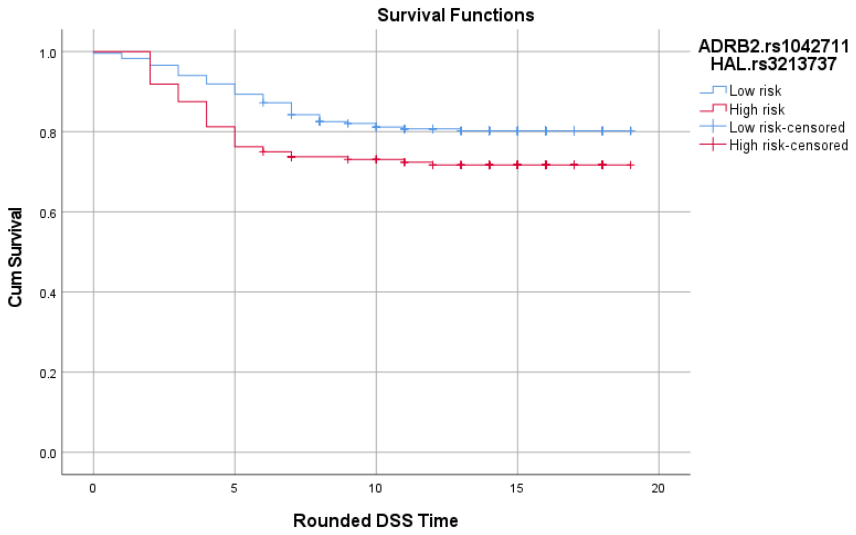


Log-rank p = 0.133385031290691

Red: AA and GG

Blue: GA

2-way model:



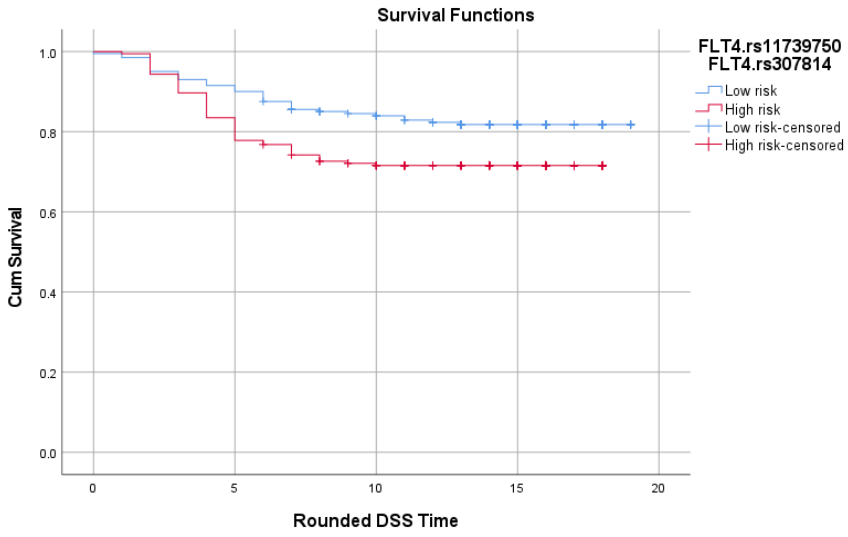
Log-rank p = 0.0309353056998482

Red: (TT,CT), (CT,TT), (CC,CT)

Blue: All other genotype combinations

VEGFC

2-way model:

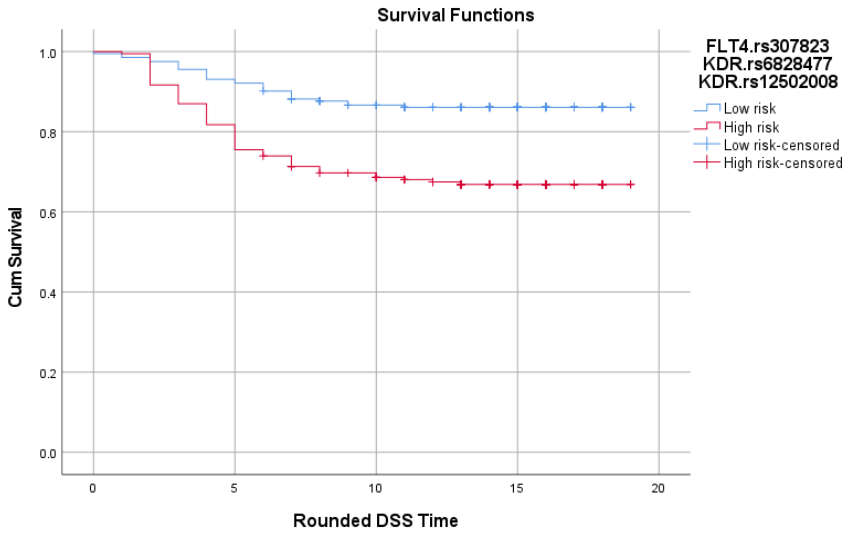


Log-rank p = 0.0126227526320323

Red: (CC,TC), (TC,CC), (TC,TT), (TT,TC)

Blue: All other genotypes except (TT, TT)

3-way model:



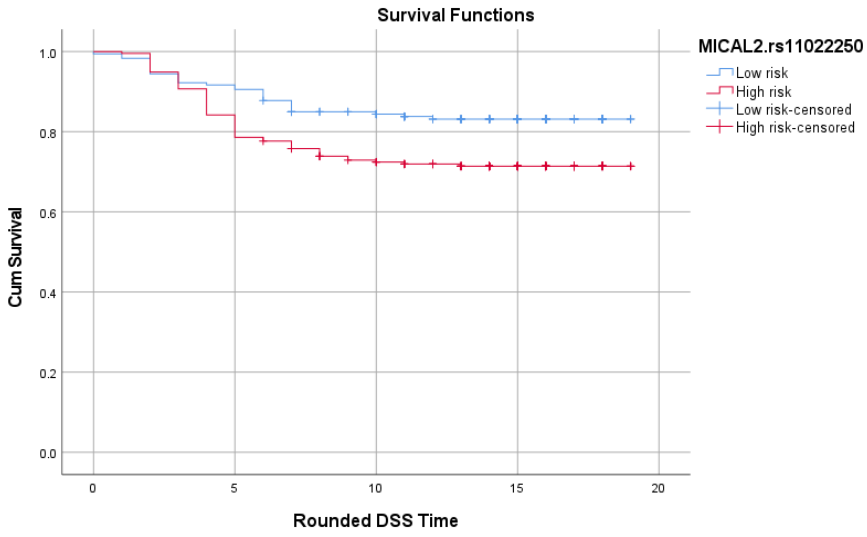
Log-rank p = 5.89522330356229E-06

Red: (AA,TT,TG), (AA,CT,GG), (AA,CC,GG), (AA,CC,TG), (GA,TT,GG), (GA,CT,TG), (GA,CC,TT), (GG,TT,TG), (GG,CT,TG), (GG,CT,TT)

Blue: All other genotype combinations except (GG,CC,TG) and (GG,CC,TT)

VEGFR1

1-way model:

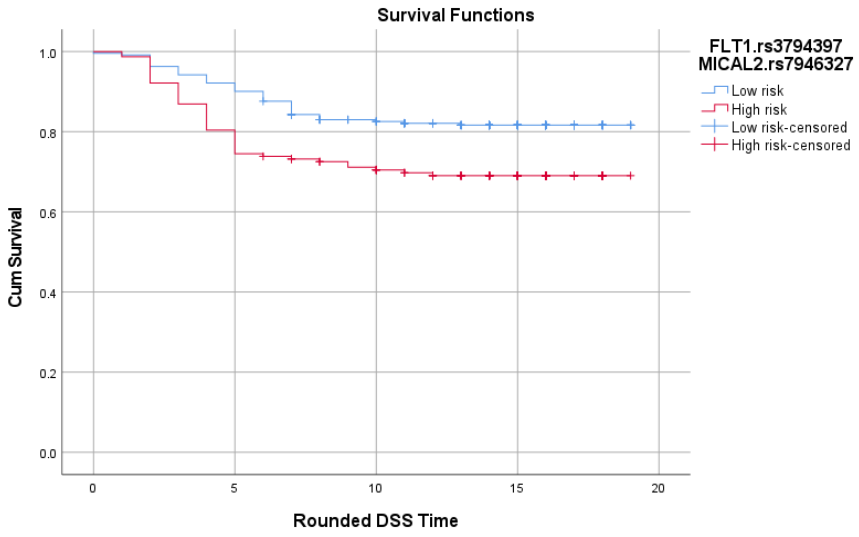


Log-rank p = 0.00694530789778492

Red: TT

Blue: GT, GG

2-way model:

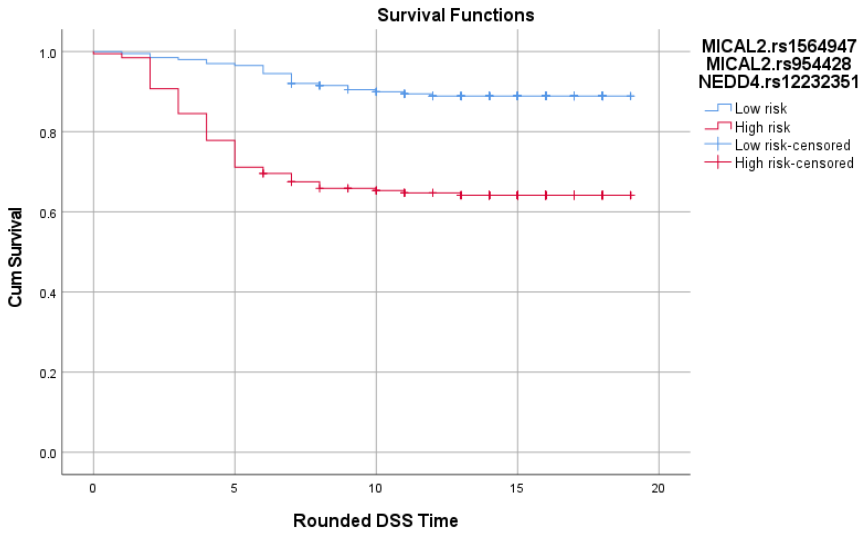


Log-rank p = 0.00244383578765834

Red: (CC,AA), (TC,CA), (TT,CA)

Blue: All other genotype combinations

3-way model:



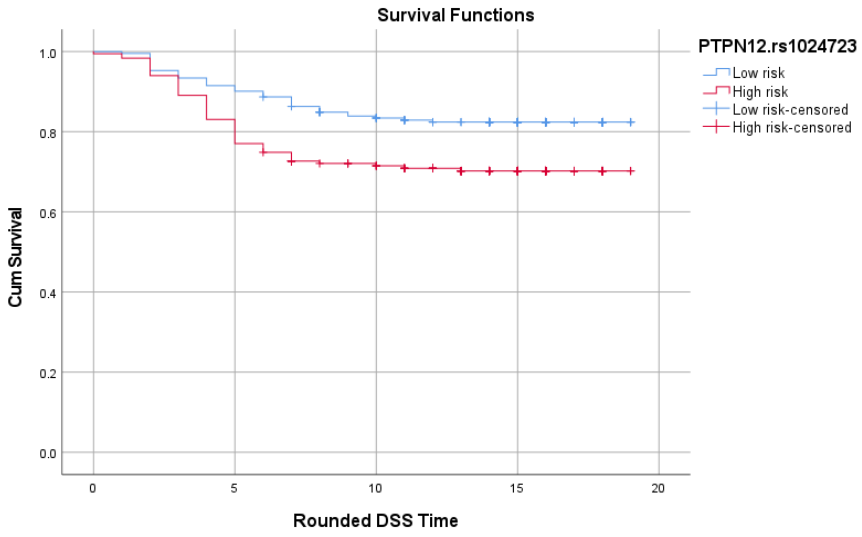
Log-rank p = 1.50474809470499E-09

Red: (GG,AA,TT), (GG,AA,AA), (GG,GA,TT), (GG,GA,AT), (GG,GG,TT), (AG,AA,TT), (AG,GA,AT), (AA,GG,TT), (AA,GG,AA)

Blue: All other genotype combinations except (AA,AA,TT) and (AA,AA,AA)

VEGFR2

1-way model, iteration 1:

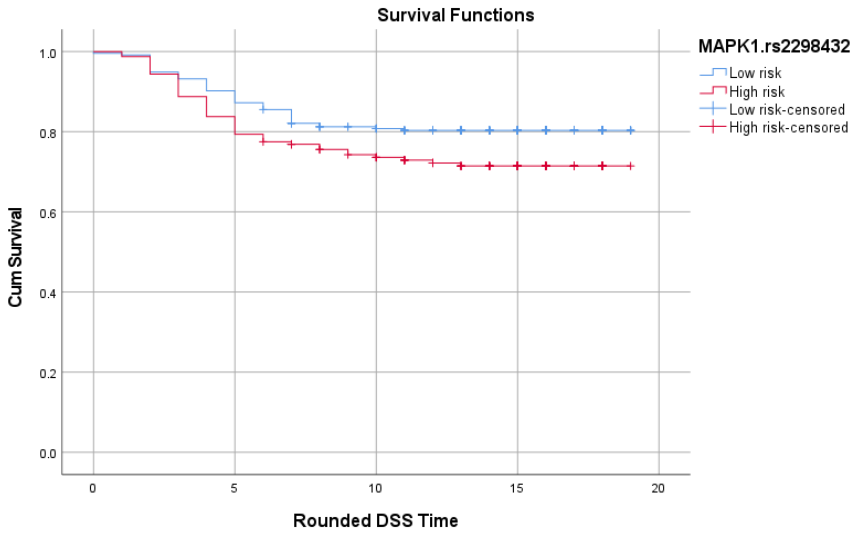


Log-rank p = 0.00325375804427782

Red: TT and CC

Blue: TC

1-way model, iteration 2:

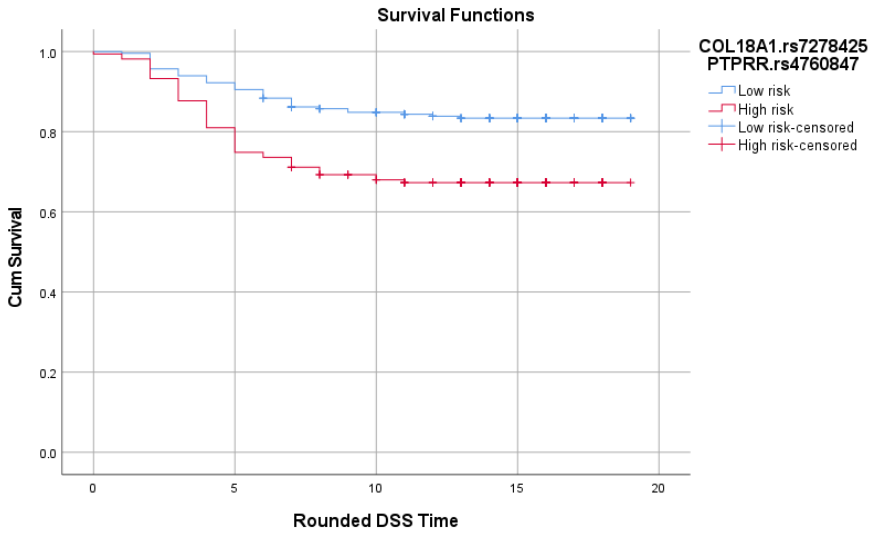


Log-rank p = 0.0440341835682624

Red: CC

Blue: AC, AA

2-way model:

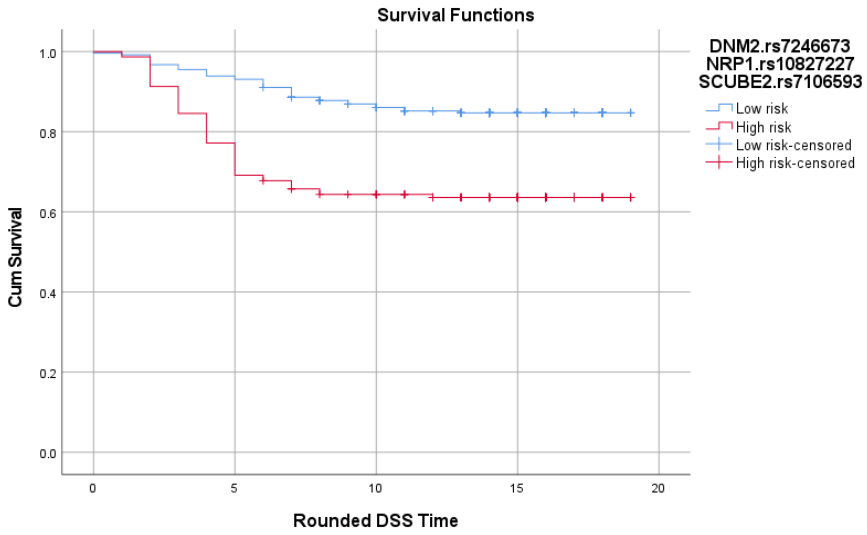


Log-rank p = 0.000128426652769784

Red: (CC,GA), (TC,AA)

Blue: All other genotype combinations

3-way model:



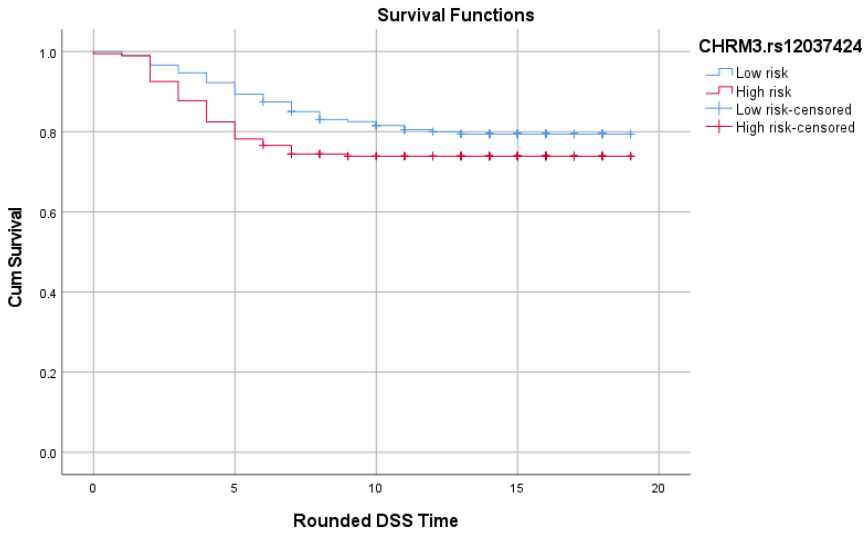
Log-rank p = 3.03572318253623E-07

Red: (TG,CC,GT), (TG,TC,TT), (TG,TC,GG), (TG,TT,TT), (TG,TT,GT), (TG,TT,GG), (TT,CC,GG), (TT,TC,GT), (TT,TC,GG), (TT,TT,GT)

Blue: All other genotype combinations

VEGFR3

1-way model, iteration 1:

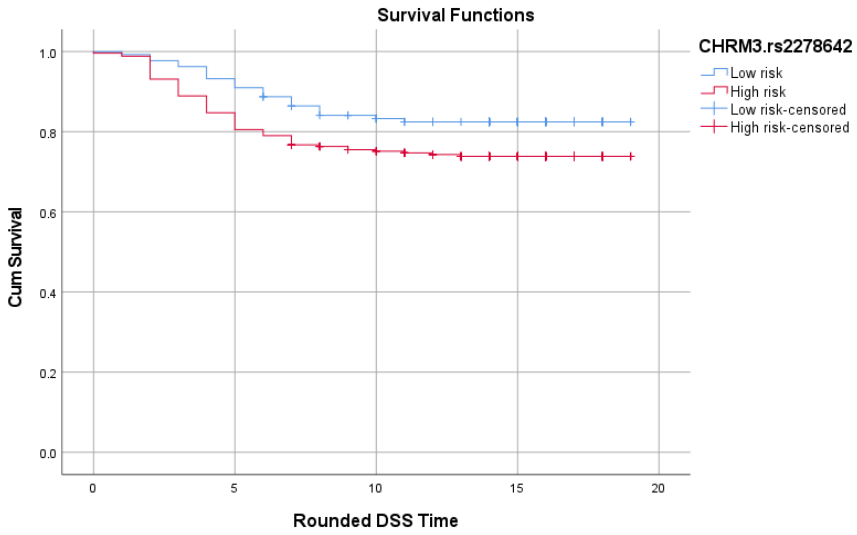


Log-rank p = 0.109869998601945

Red: TT

Blue: CT, CC

1-way model, iteration 2:

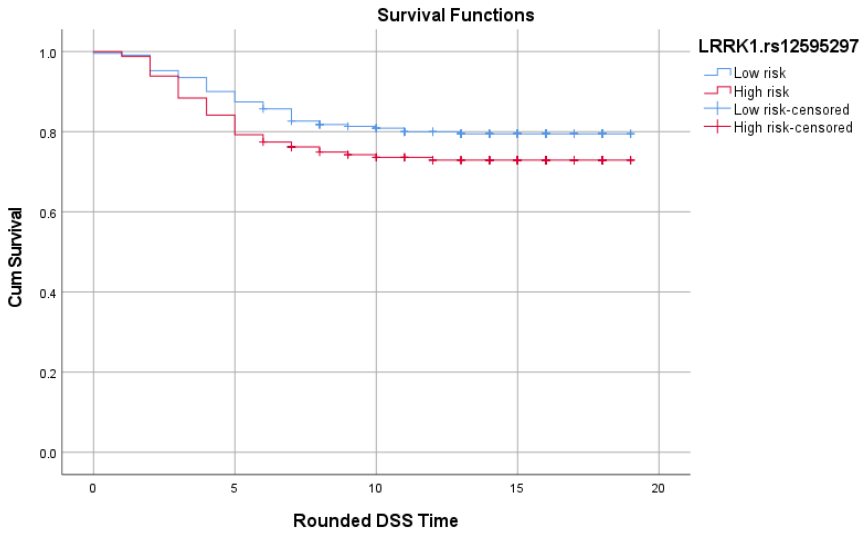


Log-rank p = 0.0443756555088326

Red: TG, TT

Blue: GG

1-way model, iteration 3:

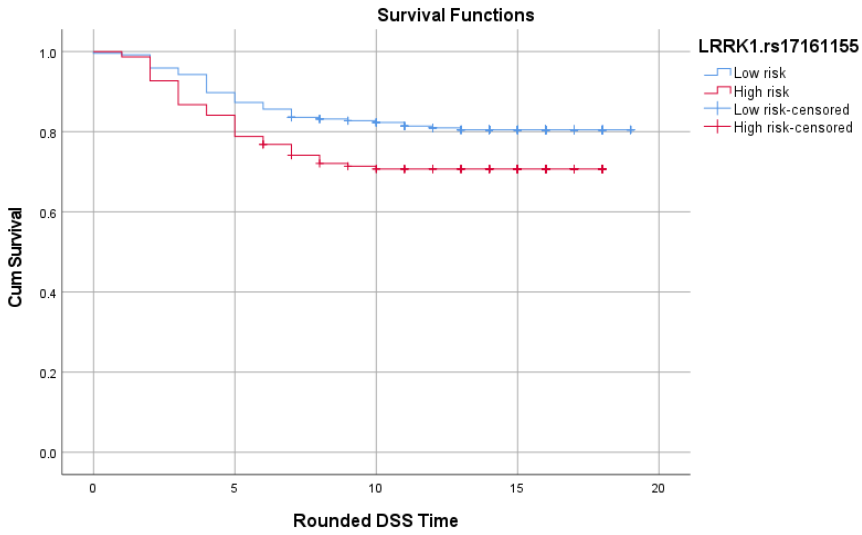


Log-rank p = 0.105405395808711

Red: GT

Blue: TT, GG

1-way model, iteration 4:

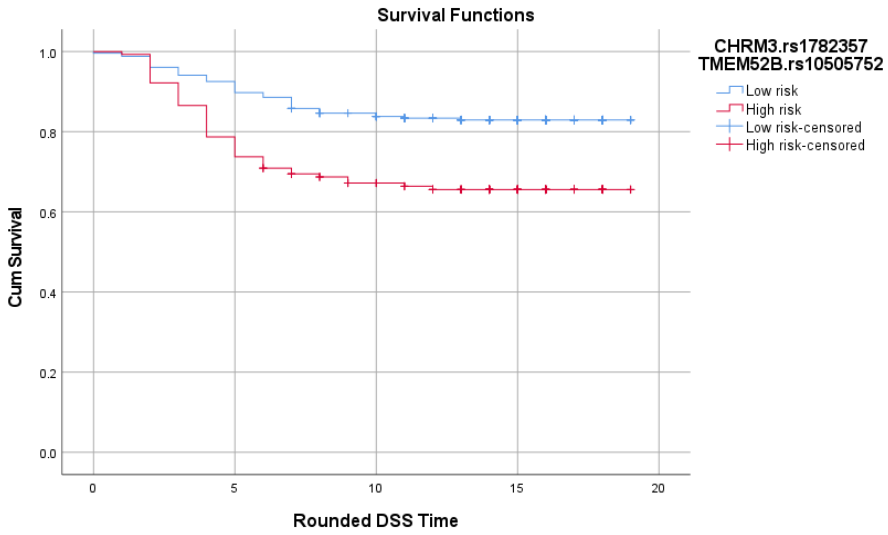


Log-rank p = 0.0190727872695036

Red: GG

Blue: AG, AA

2-way model:

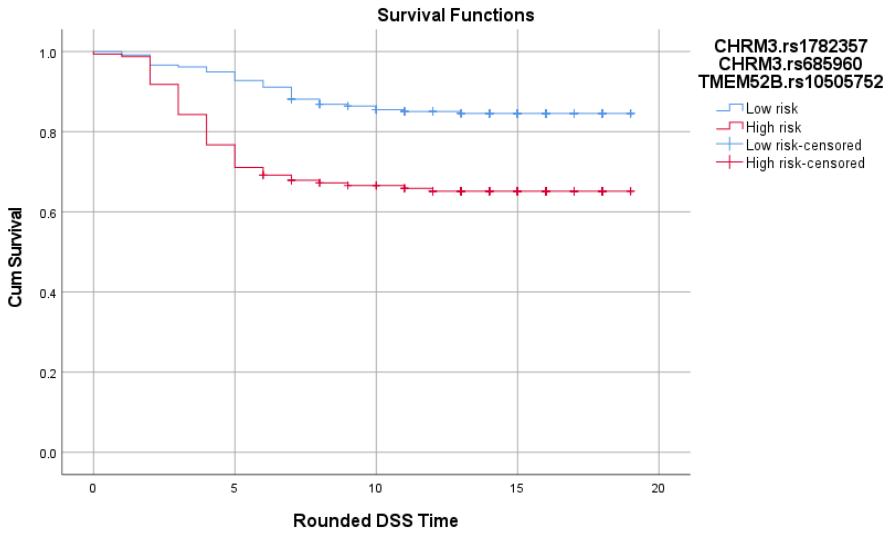


Log-rank p = 0.000050799778711075

Red: (CC,CC), (TC,TC), (TC,TT), (TT,TT)

Blue: All other genotype combinations

3-way model:



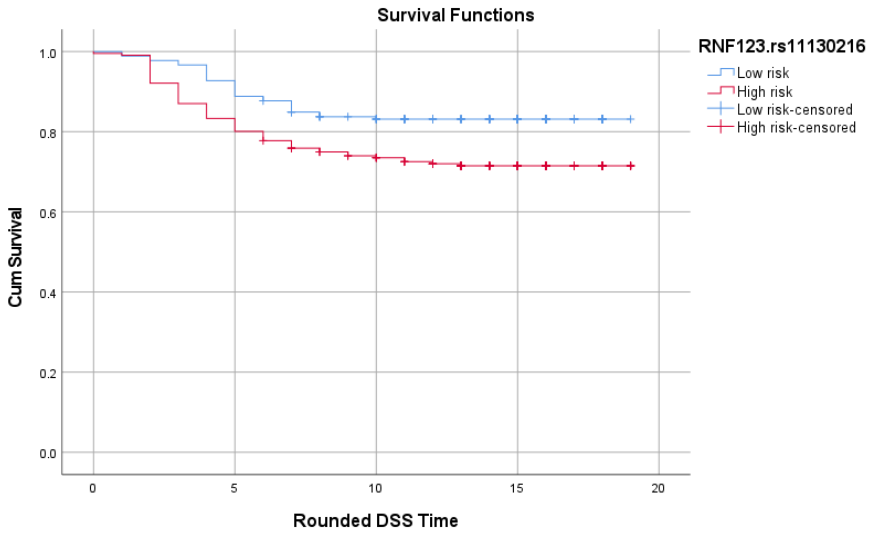
Log-rank p = 0.0000018569215896555

Red: (CC,TT,CC), (CC,CT,CC), (TC,TT,TC), (TC,TT,TT), (TC,CT,CC), (TT,TT,TT), (TT,CT,CC), (TT,CT,TT)

Blue: All other genotypes except (CC,CC,CC), (CC,CC,TC), (CC,CC,TT), (TC,CC,CC), (TC,CC,TC), (TC,CC,TT), (TT,CC,CC), and (TT,CC,TT)

PIGF

1-way model:

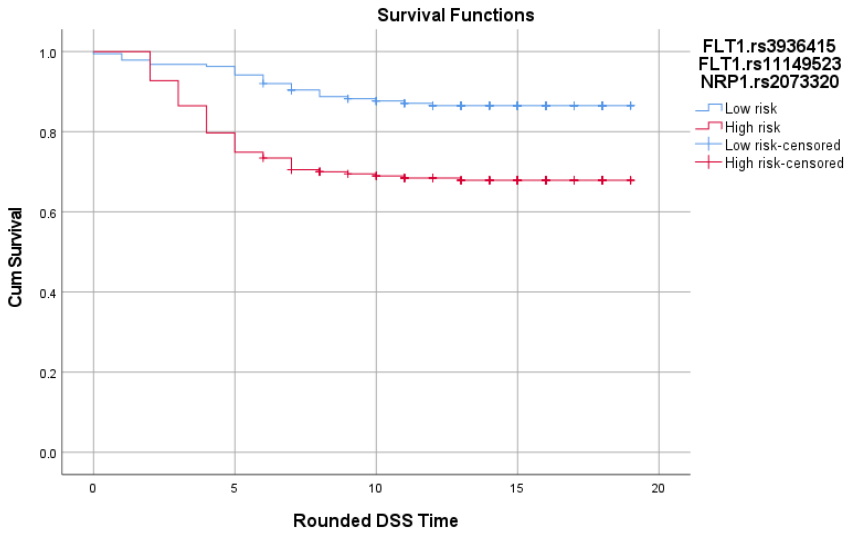


Log-rank p = 0.00581230466407173

Red: AC, AA

Blue: CC

3-way model:



Log-rank p = 6.66585629309672E-06

Red: (GG,GG,CC), (GG,GG,TC), (GG,AG,CC), (GG,AG,TT), (GG,AA,CC), (GG,AA,TT), (AG,GG,TT), (AG,AG,TC), (AG,AA,TC), (AG,AA,TT), (AA,GG,CC), (AA,AA,CC)

Blue: All other genotype combinations

Table S7. SNPs annotated as eQTLs and identified by either Cox-MDR or GMDR 0.9 in this study.

Interactor set	SNP	Gene	MAF	Regulome DB rank/score	RegulomeDB if eQTL - which tissue?	RegulomeDB if eQTL - which gene?	GTEx (if eQTL?) in transverse colon	GTEx if eQTL in transverse colon - which gene?	GTEx (if eQTL?) in sigmoid colon	GTEx (if eQTL?) in sigmoid colon - which gene?
	PART 1 (MMP family genes)									
	2-way GMDR 0.9									
	rs7817382 and rs2254207									
	rs7817382	MMP16	0.2506	6	No results	NA	cis-eQTL	MMP16 (minor allele G - lower expression)	No results	NA
	rs2254207	MMP24	0.2585	4	No results	NA	No results	NA	cis-eQTL	MMP24-AS1 (minor allele C - lower expression)

	3-way GMDR 0.9									
	rs2664369, rs11225332 and rs11639960									
	rs11639960	MMP2	0.3497	1f	monocyte (cis-eQTL)	AYTL1 (also named LPCAT2, (based on:(Zeller et al., 2010) minor allele G - lower expression)	No results	NA	No results	NA
	PART 2 (VEGF family networks)									
	1-way Cox-MDR									

PIGF [also identified by GMDR 0.9 GMDR]	rs11130216	RNF123	0.3125	4	No results	NA	cis-eQTL	RBM6 (minor allele A -lower expression), UBA7 (minor allele A -lower expression), GPX1 (minor allele A -lower expression), AMT (minor allele A - higher expression)	cis-eQTL	RBM6 (minor allele A - lower expression), AMT (minor allele A - higher expression), CCDC36 (minor allele A - higher expression), MST1R (minor

										allele A - higher expression) , CDHR4 (minor allele A - higher expression)
	1-way GMDR 0.9									
VEGFR2 (iteration 1)	rs1024723	PTPN12	0.41	5	No results	NA	cis-eQTL	APTR (minor allele T - higher expression)	cis-eQTL	APTR (minor allele T - higher expression)

PIGF	rs11130216 (iteration 1)	RNF123	0.3125	4	No results	NA	cis-eQTL	RBM6 (minor alleles A - lower expression), UBA7 (minor alleles A - lower expression), GPX1 (minor alleles A - lower expression), AMT (minor alleles A - higher expression)	cis-eQTL	RBM6 (minor allele A - lower expression), AMT (minor allele A - higher expression), CCDC36 (minor allele A - higher expression), MST1R (minor

										allele A - higher expression) , CDHR4 (minor allele A - higher expression)
VEGFR2	rs2298432 (iteration 2)	MAPK1	0.3663	3a	No results	NA	cis-eQTL	LL22NC03-86G7.1 (minor allele A - higher expression), TOP3BP1 (minor allele A - lower expression)	cis-eQTL	LL22NC03-86G7.1 (minor allele A - higher expression) , TOP3BP1 (minor

										allele A - lower expression) , PPIL2 (minor allele A - higher expression)
VEGFR3	rs17161155 (iteration 4)	LRRK1	0.3887	1f	monocyte (cis-eQTL)	LRRK1 (based on (Zeller et al., 2010), minor allele A - lower expression)	No results	NA	No results	NA
	2-way GMDR 0.9									
VEGFA	ELAVL1.rs3786619 FLT1.rs3936415									

	rs3786619	ELAVL1	0.47	4	No results	NA	cis-eQTL	CTD-3193O13.8 (minor allele A - lower expression)	cis-eQTL	CTD-3193O13.8 (minor allele A - lower expression), CTD-2325M2.1 (minor allele A - higher expression)
VEGFB	ADRB2.rs1042711 HAL.rs3213737									
	rs3213737	HAL	0.4088	5	No results	NA	cis-eQTL	AMDHD1 (minor allele A - lower expression)	cis-eQTL	AMDHD1 (minor allele A - lower expression)

VEGFC	FLT4.rs11739750 FLT4.rs307814									
	rs11739750	FLT4	0.2175	1f	monocyte (cis-eQTL)	SCGB3A1 (based on(Zeller et al., 2010): minor allele T - higher expression)	No results	NA	No results	NA
	3-way GMDR 0.9									
VEGFA	CLU.rs9331888 ELAVL1.rs3786619 NRP2.rs861079									
	rs3786619	ELAVL1	0.47	4	No results	NA	cis-eQTL	CTD-3193O13.8 (minor allele A - lower expression)	cis-eQTL	CTD-3193O13.8 (minor allele A - lower expression) ,

										CTD-2325M2.1 (minor allele A - higher expression)
VEGFB	ADRB2.rs1042711 NRP1.rs17296436 VEGFB.rs11603042									
	rs11603042	VEGFB	0.36	5	No results	NA	cis-eQTL	TRPT1 (minor allele T - higher expression), FKBP2 (minor allele T - higher expression)	cis-eQTL	TRPT1 (minor allele T - higher expression), FKBP2 (minor allele T - higher expression)

VEGFC	FLT4.rs307823 KDR.rs6828477 KDR.rs12502008									
	rs12502008	KDR	0.3613	4	No results	NA	No results	NA	cis-eQTL	SRD5A3 (minor allele T - higher expression)
VEGFR1	MICAL2.rs1564947 MICAL2.rs954428 NEDD4.rs12232351									
	rs12232351	NEDD4	0.335	6	No results	NA	cis-eQTL	NEDD4 (minor allele A - higher expression)	No results	NA
VEGFR2	DNM2.rs7246673 NRP1.rs10827227 SCUBE2.rs7106593									

	rs7106593	SCUBE2	0.425	7	No results	NA	No results	NA	cis-eQTL	TRIM66 (minor allele G - lower expression), SCUBE2 (minor allele G - higher expression)

Only the variants that are annotated as an eQTL are shown in this table. eQTL: expression quantitative trait locus; MAF: Minor Allele Frequency; NA: not applicable

Appendix 2: Ethics approval

Researcher Portal File #: 20182055

Dear Mr. Aaron Curtis:

This e-mail serves as notification that your ethics renewal for study HREB # 2018.051 – Examining interactions among different variables that can explain the prognostic variability in colorectal cancer – has been **approved**. Please log in to the Researcher Portal to view the approved event.

Ethics approval for this project has been granted for a period of twelve months effective from July 17, 2022 to July 17, 2023.

Please note, it is the responsibility of the Principal Investigator (PI) to ensure that the Ethics Renewal form is submitted prior to the renewal date each year. Though the Research Ethics Office makes every effort to remind the PI of this responsibility, the PI may not receive a reminder. The Ethics Renewal form can be found on the Researcher Portal as an "Event".

The ethics renewal [will be reported] to the Health Research Ethics Board at their meeting dated August 25, 2022.

Thank you,

Research Ethics Office

(e) info@hrea.ca
(t) 709-777-6974
(f) 709-777-8776
(w) www.hrea.ca
Office Hours: 8:30 a.m. – 4:30 p.m. (NL TIME) Monday-Friday