GENOME-WIDE ASSOCIATION STUDIES INVESTIGATING A HISTOLOGICAL VARIANT AND TIME-TO-METASTASIS OF COLORECTAL CANCER

By

© Michelle Penney

A thesis submitted to School of Graduate Studies

in partial fulfillment of the requirement for the degree of

Master of Science in Medicine (Human Genetics)

Discipline of Genetics/Faculty of Medicine

Memorial University of Newfoundland

MAY 2018

St. John's Newfoundland

ABSTRACT

Colorectal cancer is a common and complex disease with significant impact on patients and their families. Despite the extensive research conducted on this disease, there is still significant variability in tumor characteristics and disease outcomes. The unknown variability may be explained, in part, by germline genetic variations. This dissertation aimed to identify genetic polymorphisms associated with colorectal cancer tumor histology as well as with the long-term risk and/or timing of metastasis in colorectal cancer using appropriate study designs and statistical methods. As a result of these comprehensive analyses, we identified a set of polymorphisms that significantly increase the discriminatory accuracy of a model for distinguishing between mucinous and non-mucinous colorectal tumors. In addition, we identified ten polymorphisms significantly associated with time-to-metastasis of colorectal cancer after adjusting for significant baseline characteristics. Once replicated, these results could assist in better understanding the complex biological mechanisms behind colorectal tumor histology and distant metastasis.

ACKNOWLEDGEMENT

This graduate program has been one of the most challenging and rewarding endeavors I have ever undertaken. First and foremost, I would like to express my utmost appreciation and gratitude to my primary supervisor, Dr. Yildiz Yilmaz. Her direction, patience, and advice have been astounding and essential in the completion of this thesis. Secondly, I don’t think that I can put into words the tremendous amount of inspiration and encouragement given to me by my co-supervisor, Dr. Sevtap Savas. The time, effort, and sacrifices she made in the journey to complete my projects were limitless. The support and guidance from both Dr. Yilmaz and Dr. Savas cannot be understated and I am eternally grateful to have had the two best mentors possible. They have instilled in me a level of confidence that I never thought possible, making me into the best version of myself and, for that, I can never thank them enough.

I would also like to take this moment to thank my committee member, Dr. Kathy Hodgkinson, for her valuable time and comments in the committee meetings as well as during the review of this dissertation. Her input has expanded my view, forcing me to consider things I never would have and thereby truly enriching my academic experience.

To the patients, personnel, and families of the NFCCR, I cannot begin to express my gratitude. In addition, I would like to thank the numerous colleagues and faculty within the Discipline of Genetics. This is such a close, welcoming, supportive group of individuals and I would have never completed this program without them.

I would also like to graciously thank the several funding agencies that made this dissertation possible. Specifically, I would like to thank TPMI/NL SUPPORT, the Discipline of Genetics, and the Neary family for their financial support. In addition, I would like to thank the funding agencies that have granted their support to my supervisor: Research & Development Corporation of Newfoundland and Labrador (RDC), the Faculty of Medicine, and Natural Sciences and Engineering Research Council of Canada (NSERC).

Finally, this newest accomplishment would have never been possible without the love and support of my family. To my parents, Barbara and Kalvin Penney, I thank you for your unwavering support, infinite confidence, and imparting me with the strength to conquer anything I put my mind to. I would also like to thank my twin brother, Matthew Penney, for his immense confidence and support, as well as helping explain my accomplishments to our parents when I couldn’t find the words. The encouragement and pride continually expressed by my brother, Andrew Flynn, and sister, Amanda Rose, touched and motivated me more than they will ever know. Lastly, I would like to thank my husband, Chris Di Battista, whose love, support, reassurance, and patience mean the world to me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABBREVIATIONS

#

5-FU	5- fluorouracil

A

AIC	Akaike information criterion
AJCC	American Joint Committee on Cancer
AMY1	Amylase 1
APC	Adenomatosis polyposis coli
AUC	Area under the curve

B

BMPR1A	Bone morphogenetic protein receptor type 1A
BRAF	B-Raf proto-oncogene

C

CASP6	Caspase 6
CCDC109B	Coiled-coil domain containing 109B
CCDC141	Coiled-coil domain containing 141
CDC42P4	Cell division cycle 42 pseudogene 4
CECR2	Cat eye syndrome chromosome region, candidate 2
CECR3	Cat eye syndrome chromosome region, candidate 3
Chr	Chromosome
CI	Confidence interval
CIHR	Canadian Institute of Health Research
CIMP	CpG island methylator phenotype
CIN	Chromosomal instability
CNV	Copy number variant

CMS	Consensus molecular subtypes
D	
DHRS4	Dehydrogenase/reductase 4
DHRS4-AS1	DHRS4 antisense RNA 1
DHRS4L2	Dehydrogenase/reductase 4 like 2
DNA	Deoxyribonucleic acid
E	
EGFR	Epidermal growth factor receptor
EMT	Epithelial-to-mesenchymal
EPHB1	Ephrin receptor B1
eQTL	Expression quantitative trait locus
F	
FAIM3	Fas apoptotic inhibitory molecule 3
FAM87A	Family with sequence similarity 87 member A
FAP	Familial adenomatous polyposis
FCMR	Fc fragment of IgM receptor
FDA	Food and Drug Administration
FHIT	Fragile histidine triad
FIT	Fecal immunochemical test
G	
GDP	Guanosine diphosphate
gFOBT	Guaiac-based fecal occult blood test
GI	Gastrointestinal
GSK3β	Glycogen synthase kinase 3 beta
GTP	Guanosine triphosphate

GWAS	Genome-wide association study

H

HIGD1AP14	HIG1 hypoxia inducible domain family member 1A pseudogene 14
HNPCC	Hereditary non-polyposis colorectal cancer
HR	Hazard ratio
HREB	Health Research Ethics Board

I

Indel	Insertion/deletion
IL21	Interleukin 21
IL24	Interleukin 24

J

JPS	Juvenile polyposis syndrome

K

KIF16B	Kinesin family member 16B
KRAS	Kirsten rat scarcoma viral oncogene

L

LD	Linkage disequilibrium
LINC00596	Long intergenic non-protein coding RNA 596
LOH	Loss of heterozygosity

M

MACROD2	MACRO domain containing 2
MAF	Minor allele frequency
MAP	MUTYH-associated polyposis

MAPK	Mitogen-activated protein kinase
MIR7515	miRNA 7515
miRNA	Micro-ribonucleic acid
MLH1	mutL homolog 1
MMR	Mismatch repair
mRNA	Messenger RNA
MSH2	mutS homolog 2
MSI	Microsatellite instability
MSI-H	MSI-high
MSI-L	MSI-low
MSS	Microsatellite stable
MUC	Mucin
MUC2	Mucin 2
MUC5AC	Mucin 5AC
MUC5B	Mucin 5B
MUC6	Mucin 6
MUTYH	mutY DNA glycosylase

N

NA	Not available
NCBI	National Center for Biotechnology Information
ND	No data
NFCCR	Newfoundland Colorectal Cancer Registry
NL	Newfoundland and Labrador

O

OR	Odds ratio

P

PC-1	Plasma cell membrane glycoprotein-1
PGM1	Phosphoglucomutase 1
PIK3CA	Phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha
PJS	Peutz-Jeghers syndrome
PMS2	PMS1 homolog 2
PPIAP17	Peptidylprolyl isomerase A pseudogene 17
PPP2R4	Protein phosphatase 2 regulatory subunit 4

R

RAS	Rat sarcoma
RASGRF2	Ras protein specific guanine nucleotide releasing factor 2
RBMXP4	RNA binding motif protein, X-linked pseudogene
RDC	Research and Development Corporation
RNA	Ribonucleic acid
ROC	Receiver operating characteristic

S

SEC24B	SEC24 homolog B
SEC24B-AS1	SEC24B antisense RNA 1
SKAT	Sequence kernel association test
SLC22A16	Solute carrier family 22 member 16
SLC35F1	Solute carrier family 35 member F1
SMAD4	SMAD family member 4
SNP	Single nucleotide polymorphism
STK11	Serine/threonine kinase 11

T

TF	Transcription factor
TGF-β	Transforming growth factor beta

TNM	Tumor-node-metastasis
TP53	Tumor protein p53
TPMI	Translational and Personalized Medicine Initiative
U	
UCSC	University of California Santa Cruz
USA	United States of America
USF1	Upstream transcription factor 1
USF2	Upstream transcription factor 2
UTR	Untranslated region
W	
WHO	World Health Organization
Z	
ZBTB20	Zinc finger and BTB domain containing 20

RESEARCH OUTPUT AND AWARDS

Abstracts

Penney ME, Yilmaz YE, Green J, Parfrey PS, Savas S. An association study of single nucleotide polymorphisms with mucinous colorectal cancer: genome-wide common variant and gene-based rare variant analyses. 5th Annual Canadian Human and Statistical Genetics Meeting, Halifax, NS, Canada. Refereed conference abstract and poster presentation, April 18, 2016.

Penney ME, Yilmaz YE, Green J, Parfrey PS, Savas S. Genome-wide association analysis of time-to-metastasis of colorectal cancer based on mixture cure model. *Genetic Epidemiology* 40 (7), 656-656. Refereed conference abstract and poster presentation at the 25th Annual International Genetic Epidemiology Society Meeting in Toronto, ON, Canada, October 26, 2016.

Penney ME, Parfrey PS, Savas S, Yilmaz YE. Genome-wide association analysis of time-to-metastasis of colorectal cancer. 6th Annual Canadian Human and Statistical Genetics Meeting, Quebec City, Quebec, Canada. Refereed conference abstract and selected for platform presentation, April 24, 2017.

Publications

He Y*, **Penney ME***, Negandhi AA, Parfrey PS, Savas S, Yilmaz YE (2018) *XRCC3* Thr241Met and *TYMS* variable number tandem repeat polymorphisms are associated with time-to-metastasis of colorectal cancer. *PLoS ONE* 13(2): e0192316 (*First authors).

Penney ME, Parfrey PS, Savas S, Yilmaz YE (2018). Associations of single nucleotide polymorphisms with mucinous colorectal cancer: genome-wide common variant and gene-based rare variant analyses. (*Submitted to a peer-reviewed journal)*

Penney ME, Parfrey PS, Savas S, Yilmaz YE (2018). Genome-wide association study of time-to-metastasis of colorectal cancer. (*To be submitted*)

Travel Awards

TPMI/NL SUPPORT Travel Funding ($1,800)	October 2016
TPMI/NL SUPPORT Travel Funding ($1,450)	April 2017

Awards

Best Presenter – Discipline of Genetics Seminar Series	May 2016
Best Presentation (M.Sc. student) – Discipline of Genetics Research Forum	May 2016

Dr. Angus J Neary Genetics Scholarship April 2017

Best Presentation (M.Sc. student) co-winner – Discipline of Genetics Research Forum May 2017

Fellowship

TPMI/NL SUPPORT Educational Funding Award May 2015

- 2-year funding, $14,000 per annum

Chapter 1: Introduction

1.1 Overview of the research projects

Colorectal cancer arises from abnormal growth and proliferation of the epithelial cells in the colon or rectum. This disease is a common malignancy worldwide, but there is a noticeably higher incidence of this disease in developed counties [1]. In Canada, in particular, it is the second most common cause of cancer in both sexes as well as the second most common cause of cancer mortality in men and third in women. The province of Newfoundland and Labrador (NL) has the highest rates of incidence and mortality of all the Canadian provinces [2].

Genetic association studies facilitate the identification of genes/genomic regions that may be influential in disease formation, disease progression, response to treatment, and clinical outcomes. Typically, these studies investigate associations between genetic markers, such as single base substitutions within DNA sequences known as single nucleotide polymorphisms (SNPs), and a given disease phenotype. Recently, genome-wide approaches have gained popularity in genetic association studies since they include genetic markers throughout the genome, enabling researchers to conduct comprehensive investigations. Such approaches have had much success in identifying genetic associations with disease susceptibility and drug responses. In addition, they have been successful in identifying genetic loci associated with patient survival outcomes.

This Master's thesis describes two genome-wide association studies that investigate associations between previously generated genome-wide SNP genotype data and select clinical features of colorectal cancer. The first genome-wide association study

(GWAS) to be discussed (Chapter 2) aimed to identify common and rare germline genetic variants that are associated with tumor histology (mucinous versus non-mucinous) in a Newfoundland colorectal cancer patient cohort. The second GWAS to be discussed (Chapter 3) aimed to identify common germline genetic variations that are associated with the long-term risk and/or timing of metastasis in a Newfoundland colorectal cancer patient cohort using a suite of statistical methods. The results from both studies provide novel genetic associations with colorectal cancer-related characteristics that may be valuable for both clinical and scientific communities.

1.2 Colorectal Cancer

1.2.1. Etiology and symptoms of colorectal cancer

Malignant colorectal tumors originate from the epithelial cells in the colon or rectum. Most cases of colorectal cancer are sporadic [3], but it is estimated that up to 10% of cases are caused by highly penetrant inherited mutations [4]. The etiology of colorectal cancer is heterogeneous with many genetic and environmental factors influencing the risk of developing this disease. Known unmodifiable risk factors include age, particularly over 50 years of age; a personal history of colorectal polyps, colorectal cancer, or inflammatory bowel disease; a family history of colorectal cancer, primarily one or more first degree blood relatives diagnosed with colorectal cancer under the age of 45; and having an inherited colorectal cancer syndrome, such as familial adenomatous polyposis (FAP) or Lynch syndrome (also known as hereditary non-polyposis colorectal cancer [HNPCC]) [5]. Additionally, while their biological contributions are not well known, a number of large-scale studies have identified genetic variations that are associated with the risk of developing colorectal cancer [6]. Modifiable behavioral or environmental risk factors for colorectal cancer include smoking, alcohol consumption, being overweight or obese, lack of physical activity, and diet, namely one high in red or processed meat [5]. There are often few symptoms of colorectal cancer at early stages, resulting in the diagnosis occurring at advanced stages of the disease in the absence of screening. Possible symptoms of colorectal cancer include, but are not limited to, anemia, blood in the stool, weight loss, pain in the abdomen, fatigue, and change in bowel movements [7].

1.2.2 Incidence and mortality rates of colorectal cancer

Colorectal cancer is a common malignancy worldwide, representing the third most common cancer in males (746,300 cases) and second in females (614,300 cases) in 2012 [1]. It is estimated that over half of the cases of colorectal cancer are diagnosed in developed regions/countries [1]. Reports suggest that this trend may be due to cultural habits, more specifically a Westernized lifestyle of high caloric intake and limited physical activity. It is also possible that the screening programs or proper recording of clinical cases in developed countries may contribute to the relatively high incidence rate compared to developing countries [8].

Colorectal cancer is also a substantial contributor to worldwide cancer mortality for both males (373,600 deaths in 2012) and females (320,300 deaths in 2012) [1]. In contrast to incidence, mortality rates from colorectal cancer is higher in developing regions/countries, which may be attributed to limited access to screening and healthcare in these countries compared to developed countries. A major cause of mortality in colorectal cancer is distant metastasis, with the most common site of metastasis being the liver [9]. The risk of metastasis and death from this disease can be reduced substantially if the colorectal tumors are detected early via screening [8,10]. In fact, clinical screening has been shown by several studies to increase colorectal cancer survival rates [11,12].

In Canada, colorectal cancer is considered a significant health problem. This disease is estimated to be the second most common cancer in 2017 with over 26,000 new cases [2]. Furthermore, colorectal cancer accounts for 12% of all cancer deaths and is the second most common cause of cancer death in Canadian men and third in Canadian women [2]. The high incidence of this disease has resulted in nationwide screening initiatives, such as Screen Colons Canada [13], which allows for the identification and resection of precancerous lesions thereby reducing cancer incidence. In fact, as of 2017, all Canadian provinces are either in the process of or have implemented colorectal cancer screening programs [14]. Initial screening for asymptomatic colorectal cancer may involve looking for the presence of occult blood in the stool released by polyps or tumors. Two types of stool tests are used in Canada: the Guaiac-based fecal occult blood test (gFOBT) and the fecal immunochemical test (FIT) [15]. The gFOBT involves a chemical reaction that occurs on a paper card which detects heme in the stool. The FIT, on the other hand, uses antibodies to detect human hemoglobin in the stool. If these tests are positive for blood in the stool, imaging would likely be suggested to determine where the blood is originating. This could include colonoscopy (examining the lining of the entire colon with a light and camera), sigmoidoscopy (examining the lining of the rectum and lower colon with a light and camera), or double contrast enema (an X-ray based procedure of entire colon and rectum using a dye called barium) [15]. If these tests show an abnormality, a biopsy may be performed during a colonoscopy or sigmoidoscopy. A diagnosis then can be made by pathological examination of the biopsied specimen. The Canadian Task Force on Preventive Health Care recommends screening for low risk adults with a stool test every two years or sigmoidoscopy every 10 years starting at 50 years of age [16].

Although advances in colorectal cancer screening and treatment have resulted in decreases in both incidence and mortality, NL has the highest age-standardized rates among the Canadian provinces [2]. In addition, NL has higher than average rates of familial cases of colorectal cancer [17,18]. For those reasons, the Newfoundland and Labrador Colon Cancer Screening Program was launched in 2012 [19]. This screening program offers one of the screening tests described in the previous paragraph, an in-home FIT test, to NL residents between the ages of 50 and 74. The FIT kit is mailed to participants' place of residence every two years from the date of entry into the program. The program is geared towards individuals who are at an average risk of developing colorectal cancer, including people who have no personal or family history of colorectal cancer, have not had a colonoscopy in the previous 5 years, and have no personal history of an inflammatory bowel condition. Individuals who have a family history of colorectal cancer, on the other hand, can be referred to the Provincial Medical Genetics Program for further evaluation [19].

1.2.3. Hereditary colorectal cancer syndromes

As mentioned earlier, while the majority of the patients affected by colorectal cancer are sporadic, a small portion of cases (up to 10%) are due to inherited mutations in known genes and have well-categorized clinical presentations [4]. The main types of hereditary colorectal cancer are hereditary nonpolyposis colorectal cancer (HNPCC), familial adenomatous polyposis (FAP), MUTYH-associated polyposis (MAP), Peutz-Jeghers syndrome (PJS), and juvenile polyposis syndrome (JPS), and can be summarized as follows:

Hereditary nonpolyposis colorectal cancer (HNPCC): Lynch syndrome and familial colorectal cancer type X (FCCX) are known collectively as HNPCC. Lynch syndrome is the most common form of hereditary colorectal cancer, accounting for up to 4% of all colorectal cancer cases [4]. Lynch syndrome is inherited in an autosomal dominant manner and is caused by loss-of-function germline mutations in DNA mismatch repair (MMR) genes. This typically results in the tumors that arise in Lynch syndrome patients having high microsatellite instability (MSI) (described in Section 1.2.4.1) [20]. It is estimated that 50% of Lynch syndrome cases are caused by mutations in *MLH1*, up to 40% are caused by mutations in *MSH2*, up to 10% are caused by mutations in *MSH6*, and mutations in *PMS2* are seen in 1% of Lynch syndrome patients [21]. Patients with this syndrome are predisposed to several cancer types, with colorectal and endometrial cancer being the most frequent. The lifetime risk for colorectal cancer in Lynch syndrome patients is 50-80% [22].

Patients with FCCX, on the other hand, have a similar clinical phenotype to Lynch syndrome and meet the HNPCC diagnostic criteria [23,24], however, their tumors lack the genetic instability characteristic [25]. Specifically, there are no germline mutations detected in the DNA MMR genes in these individuals. In fact, despite extensive genetic research, the exact genetic cause of this disease remains unknown [26]. Both Lynch syndrome and FCCX have been extensively studied in the NL population [17,18].

Familial adenomatous polyposis (FAP): The second most common form of inherited colorectal cancer, FAP, has a prevalence of 1 in 10,000 [27]. FAP is an autosomal dominant disorder caused by germline mutations in the *APC* gene on chromosome 5q21 [28,29]. According to the Gene Entrez [30], this gene is a tumor suppressor gene and codes for a member of the Wnt signaling pathway. The genomic location of the causal mutation within the *APC* gene has been linked to the severity of adenomas, age of onset, and presence of extracolonic manifestations. While such correlations have been identified, there is still variability in disease features even between patients with identical mutations [31].

Patients with FAP develop hundreds to thousands of colorectal adenomatous polyps. These adenomas begin developing in early adolescence and, if left untreated, will develop into colorectal cancer by 40 to 50 years of age [32]. A less severe phenotype of this disease is attenuated FAP, characterized by a 69% average lifetime risk of developing colorectal cancer, a range of 0 to 100 adenomas, a later age of adenoma and colorectal cancer development, and a tendency towards proximal colorectal tumors [33].

MUTYH-associated polyposis (MAP): This is an autosomal recessive syndrome that causes an increased risk of developing colorectal cancer. MAP is phenotypically similar to FAP, however, patients with MAP do not have germline mutations in *APC*. MAP, instead, is caused by bi-allelic mutations in the *MUTYH* gene [34]. According to Gene Entrez database, this gene is involved in base-excision repair, a DNA repair mechanism that repairs oxidative DNA damage [30].

This syndrome is characterized by the presence of adenomatous polyposis of the colorectum [34]. The average age of diagnosis of polyposis is 43 in patients with MAP, although polyps and subsequent development of colorectal cancer can happen at earlier ages [35].

Peutz-Jeghers syndrome (PJS): PJS is a hamartomatous polyposis syndrome inherited in an autosomal dominant manner. The only known cause of PJS is germline mutations in *STK11*, a tumor suppressor gene [36,37]. The characteristic feature of PJS is two or more hamartomatous polyps in the small bowel, colon, or stomach [38,39]. The most common extracolonic manifestation of this disease is mucocutaneous pigmentation on the lips, inside the mouth, and around the eyes presenting in childhood. Gastrointestinal complications usually occur in the first decade of life and include small bowel obstruction and gastrointestinal bleeding [40].

Patients with this syndrome are at an increased lifetime risk of any cancer, with most cancers being of gastrointestinal origin [41].

Juvenile polyposis syndrome (JPS): JPS is caused by germline mutations in either *SMAD4* or *BMPR1A* [42,43]. Like PJS, JPS is also a hamartomatous polyposis syndrome inherited in an autosomal dominant fashion [44]. However, JPS does not have clear extracolonic manifestations. The main clinical feature of JPS is several juvenile polyps, mainly in the colon but can also appear in other areas of the gastrointestinal tract. Symptoms of JPS include anemia and gastrointestinal bleeding. Patients with this syndrome have an estimated lifetime risk of colorectal cancer of 39% [45], but also an increased risk of gastric, small bowel, and pancreatic cancers [46].

1.2.4 Subtypes of colorectal cancer

Colorectal cancer is a complex disease with a wide collection of factors contributing to its risk and progression [47]. This heterogeneity results in clinically and genetically distinct subtypes of colorectal cancer, including molecular subtypes and histological variants, which have noticeable differences in various aspects of the disease [48-50].

1.2.4.1 Molecular subtypes

The molecular heterogeneity of colorectal cancer is well-known and can result in differences in clinical features such as progression, therapeutic response, and outcomes. So far, many different molecular classifications have been proposed. For example, Roepman et al. (2014) proposed three subtype classifications driven primarily by their epithelial-to-mesenchymal transition (EMT) status, DNA MMR system status, and tumor proliferation rate [48]. Additionally, Budinska et al. (2013) defined five different colorectal cancer subtypes based on differential expression of 54 gene modules [50]. In fact, the presence of so many molecular subtype classifications inspired several experts to attempt to make consensus classifications [51]. However, these subtypes are still considered subjective and are not widely used, making scientific and clinical discoveries based on these subtypes challenging to generalize. Consequently, the scientific community can instead categorize patients based on the three known major molecular pathways important in colorectal cancer pathogenesis: the chromosomal instability (CIN), microsatellite instability (MSI), and CpG-island methylator phenotype (CIMP) pathways [52].

Chromosomal instability (CIN) subtype: This tumor subtype accounts for approximately 70% of sporadic colorectal cancer cases [53]. The hallmark of CIN is an aggregation of several structural genetic abnormalities, including gene deletions, duplications, and chromosomal rearrangements. As a result of such variations, these tumors typically have frequent loss of heterozygosity (LOH) of tumor suppressor genes and aneuploidy. In addition, these tumors typically have an accumulation of mutations in specific oncogenes and tumor suppressor genes, such as *KRAS*, *PIK3CA*, *APC*, and *TP53* [53]. This cancer subtype is typically indicative of poor patient prognosis [54,55].

Microsatellite instability (MSI) subtype: MSI is a form of genetic instability caused by deficiencies in the DNA MMR system, resulting in frequent variation in microsatellite sequences along the DNA [56]. MSI is detected based on the analysis of five microsatellite loci (BAT25, BAT26, D2S123, D5S346 and D17S250) in tumor genomes. Altered size of at least two of these five microsatellite panel markers is required for a tumor to be labelled as being MSI-high (MSI-H). If there is only one or no abnormal marker in the panel, the tumor is classified as MSI-low (MSI-L) and microsatellite stable (MSS), respectively [57]. In contrast to the CIN pathway, this subtype is typically diploid and LOH is rare [56].

MSI-H tumors account for approximately 15% of sporadic colorectal tumors [56]. It is mainly caused by epigenetic silencing of DNA MMR genes via promotor hypermethylation in sporadic colorectal tumors, with *MLH1* being the most common culprit (>80%) [56]. However, as previously discussed, it is also caused by inherited mutations in DNA MMR genes and is a characteristic of the tumors of patients affected by Lynch syndrome [20].

While MSI-H tumors do not respond well to 5-fluorouracil-based treatments [58], a systematic review found that the MSI-H tumor phenotype is associated with longer survival times, adverse-free disease progression, and substantially reduced instances of distant metastasis [59]. Therefore, MSI-H molecular subtype is generally recognized as a marker for favorable patient outcomes.

CpG island methylator phenotype (CIMP) subtype: The main attribute of the CIMP tumor subtype is widespread CpG island methylation [60]. This subtype is hence an example of epigenetic instability. CIMP tumors can also be MSI-H due to methylation of *MLH1*, but the two molecular events are not fully concordant [61]. There is a trend towards CIMP-positive tumors being located in the proximal colon compared to the distal colon [62]. CIMP-positive tumors can sometimes be further classified into CIMP-high, which are shown to be associated with *BRAF* mutations and MSI-H status, and CIMP-low, which are shown to be associated with *KRAS* mutations and MSS status. CIMP-negative tumors, on the other hand, are typically MSS with *TP53* mutations [63]. A systematic review and meta-analysis published in 2014 showed that the CIMP subtype is significantly associated with poor prognosis, regardless of the tumor MSI status [64].

While these molecular subtypes of colorectal cancer possess their own biological characteristics and demonstrate their own associated clinical features, interestingly, they are not necessarily mutually exclusive (as demonstrated by CIMP-positive tumors exhibiting MSI through *MLH1* promotor methylation). In contrast, histological subtypes of colorectal cancer do not display such overlap in subtype definition and, thus, can be fully distinguished from each other as explained in the next section.

1.2.4.2 Histological variants

The World Health Organization (WHO) reports a number of histological variations of colorectal cancer, which are classified based on the microscopic anatomy and characteristics of the tumor cells [65]. Since some of these histological variants are rare and have had limited investigation, I will only discuss the following subtypes: adenocarcinoma, mucinous adenocarcinoma, signet-ring cell carcinoma, and medullary carcinoma.

Adenocarcinoma: These tumors comprise the most common type of colorectal carcinoma, accounting for over 90% of cases. Adenocarcinomas form in secretory epithelial cells, which normally secrete mucus in the colon and rectum [66].

Mucinous adenocarcinoma: This is a subtype of adenocarcinomas that account for up to 15% of colorectal tumors [67]. The defining characteristic of these tumors is a high extracellular mucin component. Specifically, over 50% of the tumor volume must be composed of mucin to obtain this classification [65]. Patients with mucinous adenocarcinoma are typically younger and at an advanced stage at diagnosis [68,69]. In addition, these tumors tend to have an inferior response to broad chemotherapy treatments [70,71]. There are several differential molecular and genetic characteristics of colorectal mucinous adenocarcinoma compared to non-mucinous adenocarcinoma, which will be further discussed in Section 2.3.

Signet-ring cell carcinoma: This subtype has a defining characteristic of >50% of the tumor cells having abundant intracellular mucin causing the nucleus to be displaced to the periphery [65]. It is a rare subtype of colorectal adenocarcinoma, accounting for <1% of all cases [72]. Tumors of this subtype typically occur in younger patients, associate with lymph node metastasis, present at an advanced stage at diagnosis [73,74], and are associated with a high frequency of MSI [75]. Signet-ring cell carcinoma has also been shown to be a predictor of poorer patient outcomes [74,76].

Medullary carcinoma: This is a rare histological variant of colorectal cancer that accounts for only 5-8 cases per every 10,000 [77]. These tumors are characterized by sheets of cancerous cells with vesicular nuclei, prominent nucleoli, abundant cytoplasm, and tumor infiltrating lymphocytes [65]. Patients with this cancer subtype tend to have tumors located in the proximal colon and with high MSI [78]. In addition, these patients tend to have early stage tumors at diagnosis and have a better prognosis, including a low incidence of disease recurrence [77,78].

In summary, there are extensive molecular and histological variations in colorectal cancer. Understanding these aspects of the biology of this disease is essential in deciphering the complex pathways to pathogenesis created by these variations, as well as how they affect the patient and their clinical features. Consequently, stratifying patients based on specific disease subtype characteristics is not only challenging, but is also critical for better prognostication and clinical care. To that end, the focus of the project described in Chapter 2 is on the mucinous and non-mucinous histological variants of colorectal cancer with an aim to identify germline genetic variants associated with tumor histology.

1.2.5 Prognostic factors in colorectal cancer

Different subtypes of colorectal tumors and their respective cellular behaviors can also specify different sets of predictive and prognostic markers [79]. Prognostic factors can be clinicopathologic features as well as molecular markers. These factors aid in identifying patients that are at a higher risk for disease recurrence and/or progression and can also estimate how a patient may respond to treatment. Select Several factors have been identified or widely examined for their prognostic importance:

Tumor staging: The tumor-node-metastasis (TNM) staging is the most important prognostic indicator to date. This staging is based on the size of the tumor and/or the extent of invasion to surrounding tissue by the tumor (T) as well as metastasis to the lymph nodes (N) or to other areas of the body (M). The TNM staging as set by the American Joint Committee on Cancer (AJCC) is the most widely used staging method [80]. The latest stage grouping was published in 2010 and is given in Table 1.1.

Lymphatic and vascular invasion: There are two ways in which tumors cells can spread to other parts of the body: via invasion into the lymphatic system (lymphatic invasion) or the blood circulatory system (vascular invasion) [81]. Since metastasis is a major cause of mortality in colorectal cancer, it is not surprising that lymphatic invasion and vascular invasion strongly correlate with cancer recurrence and survival. In fact, it has been estimated that the 5-year survival rate drops to 30-35% in the event of lymphatic invasion or vascular invasion, where

Table 1.1 AJCC staging for colorectal cancer

Stage	Tumor	Node	Metastasis
Stage 0	Tis	N0	M0
Stage I	T1, T2	N0	M0
Stage II	T3, T4	N0	M0
Stage IIA	T3	N0	M0
Stage IIB	T4a	N0	M0
Stage IIC	T4b	N0	M0
Stage III	Any T	N1, N2	M0
Stage IIIA	T1, T2	N1	M0
	T1	N2a	M0
Stage IIIB	T3, T4a	N1	M0
	T2, T3	N2a	M0
	T1, T2	N2b	M0
Stage IIIC	T4a	N2a	M0
	T3, T4a	N2b	M0
	T4b	N1, N2	M0
Stage IVA	Any T	Any N	M1a
Stage IVB	Any T	Any N	M1b

Tis: carcinoma in situ; T1: tumor invades submucosa; T2: tumor invades muscularis propria; T3: tumor invades subserosa or beyond; T4: tumor pierces visceral peritoneum and/or directly invades other organs or structures; T4a: tumor pierces visceral peritoneum; T4b: tumor directly invades other organs or structures; N0: no regional lymph node metastasis; N1: metastasis to 1-3 lymph nodes; N1a: metastasis to 1 lymph node; N1b: metastasis to 2-3 lymph nodes; N1c: tumor deposits in the subserosa or beyond without lymph node metastasis; N2: metastasis to 4 or more lymph nodes; N2a: metastasis to 4-6 lymph nodes; N2b: metastasis to 7 or more lymph nodes; M0: no distant metastasis; M1a: distant metastasis to one organ; M1b: metastasis to more than one organ or site.
Used with permission from the AJCC Cancer Staging Handbook, 7th Edition (2010) [80] (Appendix A).

this rate is up to 84% when there is no evidence of invasion [81].

Tumor budding: Tumor budding denotes microscopic clusters of de-differentiated cancer cells at the invasive front of the tumor. This characteristic is an indicator of aggressive tumor behavior that may be driven by an EMT-like process [82]. Several studies have shown that severe tumor budding correlates with increased risk of disease recurrence and reduced survival rates [83]. Particularly in stage II patients, it has been generally observed that patients who experience highly adverse clinical outcomes are those with evidence of tumor budding [84]. Hence, tumor budding may have significant prognostic potential, particularly in predicting aggressive tumor behavior and metastasis.

Tumor grade: Tumor grade is an account of how far the tumor cells and tissue have deviated from normal cells and tissues [85]. A high grade tumor is indicative of tumors that are undifferentiated or poorly differentiated and look extremely abnormal under a microscope when compared to their normal counterparts. Conversely, low grade tumors are well-differentiated and look similar to their normal counterparts. High grade tumors tend to associate with a less favorable patient prognosis and high metastatic potential whereas low grade tumors tend to infer a favorable patient prognosis and low metastatic potential [85]. However, some experts have proposed that a lack of a consensus grading scheme makes the prognostic evaluation of tumor grade difficult [86,87].

Tumor location: Traditionally, it has been shown that tumors located in the rectum tend to have worse survival compared to colonic tumors [88,89]. This is mainly because of the substantially higher risk of local recurrence in rectal cancer patients compared to colon cancer patients [90]. This trend appears to be changing as some studies showed that advances in treatment are related to improved survival in stage II-III colorectal cancer patients regardless of tumor location [91,92]. However, this improvement in survival is not yet universal and it was suggested in a recent study that the location of the tumor still correlates with clinical outcomes [93].

Demographic and life-style related variables: Several demographic variables have been evaluated for their prognostic significance in colorectal cancer. An increased age at diagnosis has been shown to be a negative prognostic factor, even though younger patients typically present with more locally advanced cancers [94,95]. In addition, several studies have identified a survival advantage in females [96-98]. Clearly, age at diagnosis and sex appear to play a role in patient prognosis. However, differential comorbidities stemming from age at diagnosis or sex offer an additional challenge in establishing a direct causal link between these characteristics and patient prognosis [80]. Last but not least, there are a number of other environmental and lifestyle related factors that have been suggested as potential prognostic variables, such as cigarette smoking [99] and lack of physical activity [100,101], which demonstrates the potential importance of environment in prognostication of this disease.

Tumor mutations: Like in other cancers, scientists have identified several recurrent somatic mutations in colorectal tumors [102]. Among these are the mutually exclusive mutations in two members of the RAS/MAPK pathway: *KRAS* and *BRAF*. This pathway controls several important cellular functions, including the regulation of cellular proliferation, differentiation, and apoptosis [103]. *KRAS* mutations are typically activating missense mutations, of which 90% are in exon 2, and can be found in 30-50% of colorectal tumors [104]. These mutations are increasingly screened for in clinics since they confer resistance to targeted treatments, including anti-epidermal growth factor receptor (EGFR) antibodies which have recently become approved by the Food and Drug Administration (FDA) for treatment of metastatic colorectal cancer [105]. *BRAF* mutations are less common, occurring in 5-10% of colorectal tumors [104]. One of the mutations identified in *BRAF* is a missense mutation, Val600Glu, that causes the gene to be perpetually activated and, thus, work as an oncogene by promoting proliferation and reducing apoptosis [106]. While their clinical utility remains to be fully established, in some studies mutations in these genes have been associated with adverse outcomes in colorectal cancer patients [107-109].

Polymorphisms and other genetic variations: Numerous studies have been conducted testing associations between genetic polymorphism and colorectal cancer patient outcomes. These associations have identified potential biomarkers for not only survival outcomes [110-112], but also responses to treatment [113-115]. However, a common challenge with genetic associations is the inconsistency among different studies or lack of replication of the results [116]. Consequently, for the time being, the results from such studies are not used in clinical care of colorectal cancer patients.

In summary, there are several variables that may influence or predict the prognosis of colorectal cancer patients. However, there is still much to discover considering the variability in outcomes experienced by the patients. Consequently, the research project described in Chapter 3 focuses on identifying germline genetic polymorphisms that are associated with the long-term risk and timing of metastasis in a colorectal cancer patient cohort from NL.

1.3 Genetic variations in humans

Genetic variations are changes in DNA that can range from single nucleotide changes to large scale structural changes. If these variations exist in the gametes, they are known as germline variations and can be passed onto future generations. On the other hand, if a variation occurs in other tissues during development, it is known as a somatic variant and is not passed onto future generations [117]. At the genetic level, all human traits may be affected by one or more genetic variations.

The most common form of genetic variation in humans is the substitution of a single base in a DNA sequence known as a SNP [118]. SNPs can occur within the protein coding sequences of genes. In these cases, SNPs are known as synonymous if there are no changes to the amino acid sequence, missense when the substitution changes the codon to encode a different amino acid, and nonsense if the change in sequence results in a stop codon and, consequently, a possibly truncated protein. In addition, SNPs can occur in regulatory regions, such as untranslated regions (UTRs) or splice sites, as well as intergenic sequences [119].

The 1000 Genomes Project and other large-scale genome projects have identified over 88 million human SNPs with allele frequencies of >1% [120]. The distribution and frequencies of human SNPs across populations show variability and can be shaped by many factors, including evolutionary mechanisms such as natural selection [121,122] and genetic drift [123]. A large portion of SNPs are common among several different populations [124], however, in some cases their frequencies can noticeably vary [120]. For example, the Q allele of the K121Q polymorphism in the glycoprotein gene *PC-1* was shown to be highly common in African American children (allele frequency up to 80%), while the allele frequency was only up to 15% in Caucasian children [125]. Overall, African populations have a larger number of genetic variation when compared to other historical human populations, namely Asians and Caucasians [120]. Additionally, in some cases, differences in allele frequencies explain sub-population variances in the incidence of traits or diseases. For example, there is a measurable difference in the frequencies of alleles associated with height between Northern and Southern Europeans that is consistent with polygenic adaptation, or weak widespread natural selection [126]. Another study demonstrated evidence that individuals of Puerto Rican ancestry have a higher frequency of risk alleles and lower frequency of protective alleles for 101 disease-associated SNPs compared to individuals of a non-Hispanic Caucasian ancestry [127]. These examples underlie the potential biological effects of genetic variations and importance of considering the differences in allele frequencies between populations when examining the genetic basis of traits.

While they are typically considered benign, a portion of SNPs can directly affect a gene and its expression or function. For example, one study revealed that a polymorphism in the gene encoding interleukin 21 (*IL21*) was associated with significant changes in IL21 mRNA and protein levels [128]. The authors of this study suggest that this upregulation of gene expression is a possible mechanism for an increased risk of ischemic stroke. Another study showed that two missense variants that affect the Asp263 residue of *PGM1* causing a decrease in catalytic activity of the encoded enzyme, which may be a cause of PGM1 deficiency in individuals with these variants [129]. As will be discussed later, identifying SNPs that have causal links to human phenotypes has been an intense research area in genetics.

In addition to SNPs, DNA variations involving more than one base pair do exist in the human genome. Two common types are insertion/deletion variants (indels) and copy number variants (CNVs). In these types of genetic variations, sequences of DNA are inserted/amplified or deleted. If the length of the affected sequence is less than 1,000 base pairs, some researchers define them as an indel. In contrast, if the length of the sequence is greater than 1,000 base pairs, it may be called a CNV [130]. Intuitively, such structural variants could have a large effect on the gene or protein function [131]. In fact, there are good examples of such functional variations in literature. Falchi et al (2014) demonstrated that a CNV encompassing the salivary amylase gene *AMY1* associates with body mass index and risk of obesity [132]. Individuals with higher copy numbers of this gene have higher amounts of the salivary amylase protein, which initiates the digestion of starch in humans. Interestingly, this CNV appears to have been positively selected for in some populations, since populations with high-starch diets have on average more copies of the *AMY1* gene, suggesting an adaptation towards more efficient starch digestion in these populations [133]. In addition, several genes related to immune responses have been shown to have CNVs with potential biological consequences [134-136]. CNVs are also known to be associated with neurological diseases, such as autism [137]. Similar to SNPs, studying indels and CNVs is an active area of research with the potential to further illuminate the influence of genetic variations on the development of complex traits.

In addition to their sizes, human genetic variations are also categorized as common or rare based on their frequencies in the populations. Typically, the categorization is based on frequency of the least common allele, also known as the minor allele frequency (MAF) [138]. The MAF threshold for common/rare classification is still subjective with different studies applying different thresholds. However, common variants typically have a MAF ≥ 1 or 5% and rare variants have a MAF < 1 or 5% [138,139]. Both common and rare variants are expected to have very important roles in complex trait development. In fact, the genetic architecture of complex traits, including diseases, can be hypothesized using the following theories: (1) the common disease-common variant hypothesis, which suggests a moderate number of common variants each contributing a moderate effect on the trait [140]; (2) the common disease-rare variant hypothesis, which advocates multiple moderate to highly penetrant rare variants with large effects are responsible for the trait [141]; (3) the infinitesimal model, postulating that a large number of small-effect common variants are responsible for complex trait development [142]; and (4) the broad-sense heritability model, which proposes a combination of gene, environment, and epigenetic interactions contribute to the trait [143]. It is likely that not all complex traits will fall into only one of these categories. As a result, it is imperative to properly investigate both common and rare variants in the exploration of the genetic basis of complex traits. These analyses require different analytical approaches, which will be discussed in Section 1.4.1.

To sum up, extensive research performed so far has shown that, while most genetic variations have little or no effect on trait development, in some cases they can have a functional influence on a gene or protein as well as biological pathways and processes. Accordingly, such variations may be the main cause of a Mendelian disease or may influence the risk of complex diseases [144].

1.3.1 Mendelian vs. complex diseases

Diseases are typically classified by geneticists as either Mendelian or complex. Mendelian disorders are usually rare but have predictable and recognizable inheritance patterns. In addition, these diseases are typically caused by high-penetrant mutation(s) in one or a few causative gene(s). Examples of such diseases are sickle-cell anemia [145] and cystic fibrosis [146]. Diverse types of genetic mutations, including point mutations, indels, and chromosomal abnormalities have been implicated in these diseases [147].

Complex diseases, on the other hand, are relatively common in the general population. These diseases can sometimes have significant genetic components, but also may require the influence of environmental factors and/or gene-environment interactions for their development. Examples of complex diseases include cancer, cardiovascular disease, and diabetes [148]. Since there are so many potential contributors to the disease, complex diseases do not typically display a predictable inheritance pattern. Consequently, the molecular biology behind complex disease etiology, including the genetic contribution, is challenging to dissect. A number of approaches can be applied to help identify genetic variations that may influence a complex phenotype, one of which is performing genetic association studies.

1.3.2 Genetic association studies

Genetic association studies can assist in the identification of susceptibility or causative alleles by testing for correlations between genetic variations and clinical outcome or complex phenotype, also known as a trait [149]. The most common genetic variations investigated in these types of studies are SNPs, however, indels and CNVs may also be considered [150].

Genetic association studies rely on detecting the causal variant either directly or indirectly. Direct detection occurs when the causal variant is directly identified through associations (i.e. genotyped and statistically associated with the phenotype). Indirect detection, on the other hand, occurs when nearby genetic markers highly correlated with the causal variant are detected to be associated with the phenotype. This non-random correlation between alleles at different genomic loci is known as linkage disequilibrium (LD). Further investigation past the association analysis is required to properly classify the variants as causal or merely a marker of the phenotype [151]. Consequently, associations with both genotyped causal SNPs and SNPs linked to the causal variant can provide valuable information related to the trait of interest and factors contributing to it.

Tests for genetic association are typically performed for each variant individually. Statistical tests are available if a researcher is interested in the simple presence or absence of an association, such as a conventional χ^2 test for association [152]. Such tests check for dependence, but do not infer a direction of effect nor can they adjust for any potential phenotype-related characteristics (i.e. exclusively univariable analyses). Alternatively, researchers may aim to explicitly model the conditional probability that a random individual in the population has the phenotype given the genetic and other variables. This is done using more complicated regression models. Regression models provide coefficients that offer a measure of association for each variable included in the model. Given that complex traits and diseases typically have several genetic and/or environmental factors influencing the development of the phenotype, regression models are desirable to model complex/multifactorial traits.

While genetic association studies have the potential to provide informative results to assist in understanding the genetic contribution to complex trait development, the human genome contains a large amount of genetic variations. Thus, data analyses can be fairly challenging. As a result, precise study designs are required to properly answer specific research questions regarding the genetic foundation of traits. Genetic association studies can focus on genetic markers in targeted genes/genomic regions (candidate polymorphism/gene/pathway studies) or throughout the human genome (genome-wide association studies; GWAS).

1.3.2.1 Candidate polymorphism/gene/pathway studies

Candidate polymorphism/gene/pathway studies are hypothesis-driven, meaning they are based on prior biological knowledge regarding a polymorphism, gene, or biological pathway relevant to the trait [149,153]. In candidate polymorphism studies, associations between a trait and individual SNPs usually with an anticipated functional link are tested. Candidate gene studies, on the other hand, test associations between several SNPs individually within a gene of expected functional importance in the development of the trait. Lastly, candidate pathway studies aim to examine variants in the genes functioning in specific biological pathways related to the etiology of the trait of interest, which may offer a more comprehensive and biologically relevant approach. While these studies aim to directly identify functional variants, as expected they may also identify markers that are linked (i.e. in high LD) with the causal functional variant within or outside the examined gene [149].

The main limitation in candidate polymorphism/gene/pathway studies is the requirement for prior knowledge thereby identifying the candidate(s). This can limit the study to a very small part of the genome. Instead, it can be useful to analyze genomic variants comprehensively in a genome-wide approach [149].

1.3.2.2 Genome-wide association studies (GWAS)

Together with the scientific knowledge created on genetic variations by large scale genome projects, such as the Human Genome Project [154] and 1000 Genomes project [120], and technological advances, such as development of DNA chips [155], high-dimensional genetic association studies using genetic markers across the genome have been made possible. In fact, GWAS have become an important contributor to genetic and complex disease research. GWAS are high-dimensional studies that are extremely helpful in identifying genetic variations implicated in complex disease risk because of their high genomic coverage, which also removes the need for selecting candidates based on prior information. Regression-based GWAS can be particularly valuable [149]. As mentioned in Section 1.3.2, regression models can analyze several variables and provide a measure of association for each variable. Consequently, regression-based GWAS can detect genome-wide genetic associations while adjusting for other disease-associated characteristics to potentially isolate the genetic effects. Using such information, researchers can develop new strategies for treatment as well as disease prevention [156-158]. Consequently, GWAS have the potential to catalyze the induction of personalized medicine [159].

While GWAS allow for a more comprehensive investigation of the genome than candidate studies, they also introduce large data processing and computational burdens that can be non-trivial challenges for researchers. In addition, since GWAS test hundreds of thousands of SNPs from the same dataset individually, there is a large burden of correction for multiple testing. These corrections can be quite conservative with such a large number of tests performed and their application can result in a loss of power of the analysis and high false-negative rates. However, these corrections are necessary to control for false-positive rates and increase the replicability of the associations [160].

The results from genetic association studies can provide important insight into the genetic foundation of complex diseases. More specifically, the results can indicate which specific genotype of a given genetic polymorphism is associated with a disease characteristic. However, the allele combinations that are investigated in such studies are dictated by the genetic model considered in the analysis, which is determined by the researcher. Accordingly, the choice of genetic model can have a strong influence on the results of the analysis.

1.3.2.3 Application of different genetic models

As mentioned, SNPs are the most common form of genetic variation and are the typical variants studied in genetic association studies. Generally, a SNP is biallelic consisting of a major and a minor allele. Consequently, a genotype can be composed of two copies of the minor or major allele (minor/major allele homozygous genotypes) or combination of both (heterozygous genotype) [161]. In different genetic models, these three possible genotypes are grouped together and compared to the other genotype(s) in different ways.

The four genetic models generally applied in genetic association analyses are the additive, dominant, recessive, and co-dominant genetic models [162]. In the additive model, patients homozygous for the minor allele are compared with the patients homozygous for the major allele at a 2x effect size and heterozygous patients are compared to the patients homozygous for the major allele at a 1x effect size. Essentially, this model assumes that the effect increases or decreases at a similar magnitude with each count of minor allele. In the dominant model, patients homozygous for the minor allele combined with heterozygous patients are compared to patients homozygous for the major allele. This genetic model assumes that one or two counts of the minor allele have similar effects (i.e. one minor allele is sufficient to produce the effect). In the recessive model, patients homozygous for the minor allele are compared with patients homozygous for the major allele combined with heterozygous patients. The recessive genetic model assumes that two minor alleles, not one, are necessary for the effect. Finally, in the co-dominant model, patients homozygous for the minor allele and heterozygous patients are compared separately to patients homozygous for the major allele. This model assumes that the effect attributable to the genotypes may be different regardless of their allele compositions. Table 1.2 illustrates these genetic models.

Table 1.2 Illustration of the different genetic models

Genetic Model	Reference Genotype	Comparison Genotype(s)	
Additive	BB	AB, AA	
Dominant	BB	AB + AA	
Recessive	BB + AB	AA	
Co-Dominant	BB	AB	AA

A: minor allele of genetic variant; B: major allele of genetic variant.

Typically, in genetic association studies, one genetic model is applied. Since study power is highest under the true/correct genetic model [163], applying an incorrect genetic model can lead to a substantial loss of power and incorrect interpretation of the results [164]. It has been suggested for genetic association studies that multiple genetic models be tested in the absence of prior biological knowledge for a GWAS [165]. Consequently, for the projects described in this dissertation (Chapter 2 and Chapter 3), we applied each of the four genetic models and tested their plausibility, as will be described in Section 1.4.3.

After the genetic models are applied to the genetic data, association analysis can be performed. It is imperative in such studies that the statistical models and methods applied are in agreement with the given data and research questions. This includes the selection of the disease outcome of interest, known as the response variable in statistical analysis. In the following sections, I will discuss different potential statistical models and methods of analysis as they relate to the response variables and specific research questions in this dissertation.

1.4 Statistical approach to the research

This thesis required extensive statistical analyses tailored to each specific research question. Different regression models were used to investigate the association between genetic markers and disease outcomes based on the type of outcome of interest. One type of disease outcome is the presence or absence of a disease or disease-specific characteristic. This is known as a binary response variable: a variable with only two possible values. These associations can provide information regarding the susceptibility of a disease or risk of developing some specific disease subtype. This type of response variable is the focus of the study presented in Chapter 2. However, if the research question is focused on disease progression, an appropriate response variable to analyze might be the time-to-event of interest. Such response variables consider the timing of a given disease-related event, such as disease recurrence or death. The results from answering these research questions, once validated, can provide information regarding disease prognosis. Such a research question is the focus of the study to be described in Chapter 3.

As mentioned in Section 1.2.4.2, the study to be described in Chapter 2 aimed to identify genetic variations that are associated with the mucinous histological variant of colorectal cancer. The presence or absence of mucinous tumor histology is a binary response variable. Consequently, we applied statistical methods that are appropriate for analyzing such data. In addition, we considered both common and rare genetic variants in the association analysis. The allele frequency has a large impact on the power of the association test and, thus, has to be a consideration in determining the appropriate statistical method to apply. As a result, I will first discuss the analysis of a binary response variable as it would be considered for investigating associations with common genetic variants, and then describe multi-marker tests for the examination of rare variants.

1.4.1 Analysis of a binary outcome

Fitting regression models can provide insight on the relationships between a binary outcome and given predictors. Binary variables take only two values, such as success/failure, yes/no, or present/absent [166,167]. These variables are often assigned 1 or 0 for the purpose of analysis. The probability of a particular outcome occurring is modelled using one or more related characteristics typically referred to as covariates, factors, or variables. Unlike the response variable, the covariates can be quantitative or qualitative and are considered potential determinants for the disease outcome. A regression model of a function of the probability of success can be used to model the association of covariates with the outcome [166,167].

Several regression models are available for the analysis of binary outcomes. However, they have their own assumptions and limitations. For example, one method to model the probability of success is using a log-linear model [168]. This model assumes a linear relationship between the logarithm of the probability of success and covariates. However, this model may yield estimated probabilities outside of the interval [0,1], which are not allowable risks. This problem can be addressed by modelling odds instead of probabilities using a logistic regression model [166,167].

1.4.1.1 Logistic regression model

When the outcome of interest is binary, the preferred method of analysis is logistic regression in many studies. Univariable logistic regression model is given as

$$log\left(\frac{p_x}{1-p_x}\right) = a + bx\ , \qquad (1)$$

where $p_x = P(D|X = x)$ is the probability of having the outcome D conditioning on a covariate $X = x$, $\frac{p_x}{1-p_x}$ is the odds of having the outcome D conditioning on a covariate $X = x$, a is the logarithm of the odds at baseline ($x = 0$), and the coefficient b is the logarithm of the odds ratio associated with a unit increase in the scale of x. Since this model calculates the logarithm of the odds for the probability of having the event, it follows that e^b is the odds ratio associated with a unit increase in the scale of x. The odds ratio is a measure of association of the given covariate x. If the odds ratio is greater than one, the comparison group has higher odds of experiencing the outcome than the reference group. If the odds ratio is less than one, the comparison group has lower odds of experiencing the event than the reference group. When the odds ratio is equal to one, the two groups have the same odds of experiencing the event. The logarithm (log) transformation ensures the estimates of p_x are between zero and one for any value of x [167,169].

In genetic association analysis, the covariate X would represent a genetic variant. Its value is determined based on the genetic model selected for each subject. Consider a SNP with minor allele A and major allele B, as in Table 1.2. The coding of each genetic model under the logistic regression model given in (1) is, as follows:

In the additive genetic model, X takes the value

$$x = \begin{cases} 2 \text{ if the genotype is } AA \\ 1 \text{ if the genotype is } AB \\ 0 \text{ if the genotype is } BB \end{cases}. \tag{2}$$

In the dominant genetic model, X takes the value

$$x = \begin{cases} 1 \text{ if the genotype is } AA \text{ or } AB \\ 0 \text{ if the genotype is } BB \end{cases}. \tag{3}$$

In the recessive genetic model, X takes the value

$$x = \begin{cases} 1 \text{ if the genotype is } AA \\ 0 \text{ if the genotype is } AB \text{ or } BB \end{cases}. \tag{4}$$

For the co-dominant genetic model, the univariable logistic regression model needs to be changed to accommodate the additional variable resulting from the consideration of each genotype separately. Consequently, when considering the co-dominant genetic model, the univariable logistic regression model becomes

$$log\left(\frac{p_{(x_1,x_2)}}{1-p_{(x_1,x_2)}}\right) = a + b_1x_1 + b_2x_2 \ , \qquad (5)$$

where

$$(x_1, x_2) = \begin{cases} (0{,}0) \text{ if the genotype is } BB \\ (1{,}0) \text{ if the genotype is } AB \\ (0{,}1) \text{ if the genotype is } AA \end{cases} . \qquad (6)$$

In the regression model (1), e^b gives the odds ratio for the different possible genotype groupings described in (2), (3), and (4). For example, consider the genotype grouping given in (3). If e^b is greater than one, individuals with the AA or AB genotypes have a higher odds of experiencing the outcome compared to individuals with the BB genotype. If e^b is less than one, individuals with the AA or AB genotypes have a lower odds of experiencing the outcome compared to individuals with the BB genotype. In the model (5), e^{b_1} provides the odds ratio for the heterozygous genotype compared to major allele homozygous genotype, and e^{b_2} provides the odds ratio for the minor allele homozygous genotype compared to major allele homozygous genotype, as denoted in (6).

In addition to the ability to provide valid risk probability measures, an asset of the logistic regression model is the ability to analyze several covariates [170]. Not only can the model handle multiple covariates, the variables do not need to be on the same measurement scale. This means some covariates could be on a continuous scale while the others can be categorical and the model can still calculate the appropriate risk measures. The multivariable logistic regression model is written as

$$log\left(\frac{p_{(x,z)}}{1-p_{(x,z)}}\right) = a + bx + c'z\,, \tag{7}$$

where $Z = \left(Z_1, Z_2, \dots, Z_J\right)'$ is a vector of J covariates other than X, $c' = \left(c_1, c_2, \dots, c_J\right)$ is a vector of coefficients for the covariates in Z, and $c'z = c_1z_1 + c_2z_2 + \cdots + c_Jz_J$ denotes a linear combination of covariate values $Z = z$ and their coefficients c. For some j ($j = 1, \dots, J$), the odds ratio for a unit increase in Z_j is e^{c_j} while holding the other covariates constant.

Now that we have defined these logistic regression models, we require methods to estimate the parameters within the models based on a random sample of the population. A well-known method to estimate the unknown coefficients in logistic regression models is the maximum likelihood estimation [167,171]. This method obtains values for the unknown parameters, such as a and b in (1), that maximize the probability of obtaining the given data. The first step in this method is to construct a likelihood function, which expresses the probability of the given data as a function of the unknown parameters under the given model. Suppose the observed data is $\{(D_i, x_i), i = 1, \dots, n)\}$ for sample size n. Since the subjects are independent of one another in a random sample, the likelihood function L is written as

$$L = \prod_{i=1}^{n} P(D_i|X = x_i)P(X = x_i)\ \ , \tag{8}$$

where $P(D_i|X = x_i)$ can be modelled by using a logistic regression model as in (1). The maximum likelihood estimators $\hat{a}$ and $\hat{b}$ of the unknown parameters a and b in model (1) are the values that maximize the likelihood function and, thus, agree the most with the given data [167,171]. Note that if the distribution of X, $P(X = x_i)$, does not depend on the unknown parameters a and b, it can be omitted from maximizing L [171]. As a result, the estimates of the unknown parameters (i.e. a and b) that maximize

$$L^* = \prod_{i=1}^{n} P(D_i | X = x_i) \tag{9}$$

are the maximum likelihood estimators (i.e. $\hat{a}$ and $\hat{b}$) of the unknown parameters. It follows that $e^{\hat{b}}$ is the odds ratio estimate for a unit increase in the scale of X.

The maximum likelihood estimation may also be used to obtain approximate confidence intervals for the odds ratio e^b and to conduct hypothesis testing for the absence of association between X and D [171]. This is because the estimators have asymptotically normal distribution when the sample size n is sufficiently large. To calculate an approximate confidence interval for the coefficient b using the normal approximation of the distribution of the maximum likelihood point estimator $\hat{b}$, we need the variance of the sampling distribution, $\hat{V}_b$ [171]. Both $\hat{b}$ and $\hat{V}_b$ are obtained from the likelihood function. Thus, an approximate $100(1 - \alpha)\%$ confidence interval can be calculated by

$$\hat{b} \pm z_\alpha \sqrt{\hat{V}_b} \tag{10}$$

where z_α is the $\left(1 - \frac{\alpha}{2}\right)$th percentile of the standard normal distribution. It is important to note that this supplies the confidence interval for b, which is the log odds ratio. To obtain the confidence interval for the odds ratio, we simply exponentiate the values calculated from the equation above. As mentioned, the likelihood function also allows for hypothesis tests to assess the absence of association between D and X. One method to accomplish this, and the test used in this thesis, is the Wald test. This method involves calculating a test statistic, $Z_b = \hat{b}/\sqrt{\hat{V}_b}$, which follows an asymptotically standard normal distribution in sufficiently large sample sizes under the null hypothesis $H_0\colon b = 0$ (i.e. no association between the covariate and outcome). In addition to the Wald test, the Score test and the Likelihood Ratio test could be used for hypothesis testing. For large n, each of these methods generally give similar p-values [171].

In this thesis, as revealed in Section 2.5.4.1, logistic regression models were fitted using R software [172]. Specifically, we used the glm function which gives the maximum likelihood estimates of the coefficients, standard errors of the maximum likelihood estimates, Wald test statistic for testing the absence of association between the outcome and covariates, and the corresponding p-values. Using the maximum likelihood estimates, their standard errors, and the asymptotic distribution assumption, we calculated the approximate confidence intervals.

Regression methods are well-known and widely used in statistical genetics and genetic epidemiology for testing single-marker genetic associations [160]. Such methods can identify genetic markers that are significantly associated with a given trait. An additional characteristic that may be of interest to researchers is how well these identified SNPs or SNP sets can differentiate between individuals who have or do not have the given trait. This can be done using receiving operating characteristic (ROC) analysis.

1.4.1.2 ROC curves for a binary outcome

A well-known method to assess the discriminatory accuracy of a given model, including various associated factors or covariates, is using a ROC curve [173]. This curve visualizes the balance between sensitivity, or true positive rate, and specificity, or true negative rate, over varying decision thresholds. These two components are inversely related, so sensitivity increases as specificity decreases. Specifically, the ROC curve plots sensitivity versus (1-specificity), which is also known as the false positive rate.

ROC curves can be constructed based on univariable and multivariable models [174]. As a result, researchers can include several associated covariates in the model and determine the change in discriminatory accuracy due to the addition of these covariates. Specifically, as covariates that are significantly associated with the binary response variable are added to the model, the discriminatory accuracy increases [175]. However, with the addition of several parameters, it is possible to overfit the model. In this case, the model becomes tailored to fit the given data instead of being able to make inferences about the population given the data [175]. Consequently, it is important to determine if there is a statistically significant increase in discriminatory accuracy when adding parameters to the model to avoid overfitting. ROC curves can also provide a direct visual comparison of several different univariable or multivariable models with differing sets of covariates on a single graph [173,176]. Such comparisons can provide some conclusions regarding the relative accuracies of the different models: a curve that lies above and to the left of another indicates a better discriminatory accuracy [176].

A useful way to quantify the discriminatory accuracy of a given model is using the area under the ROC curve (AUC). This is the combined measure of sensitivity and specificity and is a good indication of the validity of the model [173]. It is a single numeric representation of the performance of the model that is obtained by calculating the area under the ROC curve. The AUC can take any value between 0.5 and 1, inclusive. If the AUC is equal to one, the model is 100% accurate at discriminating individuals given the set of variables. This is enormously improbable in practice. Thus, an AUC of close to 1 is a more real-world objective. The minimum AUC value is 0.5, because this indicates the model has an equal chance of correctly and incorrectly discriminating individuals given a set of variables.

The interpretation of the AUC for ROC curves can be considered in the following in three ways: (1) the AUC is an average of sensitivity for all values of specificity; (2) the AUC is an average of specificity for all values of sensitivity; and (3) the AUC is the probability that a person with a given trait has a test result indicating a greater risk of that trait than a person that does not have the trait [173,176]. In medical research, ROC curves are most commonly used in testing the validity of diagnostic tests, namely in their ability to correctly identify diseased and non-diseased individuals [177,178]. For example, an AUC of 0.75 indicates that 75% of the time, a randomly selected individual with cancer has a positive cancer screening result indicating a higher risk of having cancer than a randomly selected individual that does not have cancer [176]. However, this principle can be applied to any model that attempts to differentiate between two groups depending on a given characteristic or set of variables. A quick literature review supports this, indicating that this methodology has been used in bioinformatics [179,180], clinical studies [181-183], and epidemiology [184-186].

In practice, there are typically two goals in ROC analysis: to determine if a test has more discriminatory accuracy than chance or to determine if one test outperforms another in terms of discriminatory accuracy. One way to assist in determining if there is a statistically significant difference in discriminatory accuracy is using confidence intervals of ROC curve AUC values [176]. For example, if a confidence interval for a model's AUC contains 0.5, it is acceptable to reject that model as the discriminatory accuracy is not sufficiently different than the one given by chance. However, if the confidence interval does not contain 0.5, the test discriminates between subjects with and without the trait better than chance. This concept can extend to the comparison of different models. When comparing two or more different models, if the confidence intervals for the AUC values of different models overlap, there might not be a significant difference in the discriminatory accuracy of one model in comparison to the other(s). However, if the confidence intervals do not overlap, there is a statistically significant difference and the ROC curve with the higher AUC value has a significantly higher discriminatory accuracy [176].

ROC analysis was performed in this thesis using R software [171], as discussed in Section 2.5.4.1. Specifically, we used the $pROC$ package [187]. Using this package, we were able to plot the ROC curves, and calculate the AUC with corresponding confidence intervals.

There are many ways to analyze data to identify associations between genetic variants and a binary outcome variable, and to determine the importance of these associations. The regression methods discussed above typically test the association of single genetic variants with the outcome variable, also known as single SNP analysis. However, the power of this analysis is highly dependent on the sample size. In addition, depending on the minor allele frequency, the adequate sample size for sufficient statistical power may be enormous. Consequently, these methods are most successful when used in the analysis of common variants but can be underpowered for the analysis of rare variants [188].

1.4.1.3 Rare variant analysis methods

Despite the success of single marker association tests in identifying some genetic associations with complex disease, a large portion of the genetic contribution is unknown. One theory behind this is that rare genetic variants are present that have a high effect on complex diseases [189]. Advancements in technology have facilitated the investigation of the role of rare variants in disease. In fact, rare variants have already been found to be associated with lipid traits [190,191], type 1 [192] and type 2 diabetes [193,194], hypertension [195], inflammatory bowel disease [196], sick sinus syndrome [197], and schizophrenia [198]. However, although it is easier to sequence rare variants, detecting significant associations between rare variants and complex traits can be challenging. This is due in part to the frequent use of single marker tests in rare variant analysis, which can be underpowered in detecting associations between rare genetic variants and complex diseases [188]. Consequently, several multi-marker tests have been proposed which aggregate one or more SNPs in a gene/genomic region into a single test statistic and test the association between this region and the disease trait of interest [199]. I will briefly summarize three different approaches of multi-marker tests for rare variant analysis.

Burden tests are aggregate tests that collapse information for multiple rare variants into a single genetic score [200,201]. A simple approach to this method is to count the number of minor alleles across all variants in a given set of SNPs having a rare variant. The higher the quantity of rare variants, the higher the genetic score. This method has the large assumption that all rare variants in the set of SNPs are causal and they all have the same effect on the phenotype. If this assumption is not satisfied, a substantial loss of power will occur.

Another method for testing rare variant associations is using a variance-component test [202-204]. This approach does not simply aggregate genetic information for a given set of SNPs, but it evaluates the distribution of genetic effects for the set of SNPs. Consequently, this method can aggregate both risk increasing and decreasing variants with different magnitudes of effects, including no effect, and analyze them appropriately. In fact, the major assumption of this method is that there is a mixture of variants with different effects, and the model is quite powerful in these cases. One example of such a test is the sequence kernel association test (SKAT) [204,205].

The burden and variance-component tests are both powerful methods in their own way, provided the corresponding assumptions are satisfied [199]. A variance-component method like SKAT is an attractive option since the assumptions are minimal. However, in the event that a set of SNPs does have a large number of causal rare variants that have the same direction of effect, the burden test is a more powerful method. However, it is impossible to know which method is best since the underlying genetic construction is rarely known. To address this, methods that combine burden and variance-component tests have been proposed [206-210]. One such approach is a weighted linear combination of SKAT and burden test statistics: SKAT-O [206,210]. This test has fewer assumptions than the individual tests and is an attractive choice because it is a robust test. However, it is important to note that, although SKAT-O is designed to reduce the assumptions of the individual burden and SKAT tests, the individual tests are still more powerful if their assumptions are met [199].

In this thesis, rare variant region-based testing was performed on genotype data using R software [172]. Specifically, as will be discussed in Section 2.5.4.2, we used the *SKAT-O* test in the *SKAT* package [211] which calculates the p-values for testing the association of the disease trait with each pre-determined set of SNPs in a genomic region. This package applies the additive genetic model only.

The methods described to now are appropriate to apply when the research question is focused around a binary disease trait, as is the case in the study to be described in Chapter 2. However, the study presented in Chapter 3 considers a time-to-event response variable. Specifically, as mentioned in Section 1.2.5, this study aimed to identify common germline genetic variants associated with the risk and timing of metastasis in colorectal cancer. To achieve this aim, the response variable used in this study was time-to-metastasis.

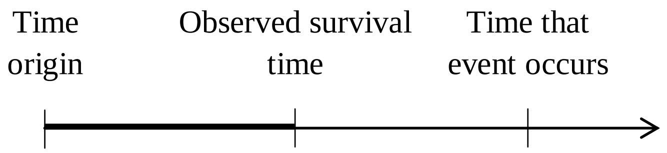
1.4.2 Analysis of time-to-event data

Prognostic studies are important contributors to medical research. Such investigations can identify both clinical and non-clinical biomarkers may confer the risk of experiencing specific disease-related outcomes. To identify such factors, researchers may wish to study survival data, typically referred to as survival analysis, which includes a time-to-event response variable. Analyzing such data requires specialized methods and considerations to appropriately describe the relationship between the timing of an event and covariates [212,213].

There are several methodological considerations of survival analysis that must be addressed prior to the beginning of the study. First, it is extremely important to have an unambiguous, clearly defined event that is conclusively measurable. In addition, the time at which the follow-up commences (i.e. the time origin) must be well-defined. Finally, the scale of measurement of time must be the same for all subjects in the study [213].

While the event definition, time origin, and time scale are exceptionally important, they are not the only considerations in the analysis of time-to-event data. The largest complication in survival analysis comes from having incomplete data, such as censored data [212,213]. Censoring is the consideration of data for unobserved events, meaning the actual time-to-event is unknown. There are different forms of censored data. The most common is right-censored data, which occurs when the true survival time is equal to or greater than the censoring time, which is the observed survival time (Figure 1.1). In fact, there are three different subsets of right censoring. Type I censoring, which occurs when

Figure 1.1 Right censored data



Event of interest occurs after the observed survival time. The observed survival time could be the end of the study or the time the participant was lost to follow-up.

there is a pre-determined follow-up time and some participants have not experienced the event by the end of this time duration. Type II censoring, on the other hand, occurs when you follow participants until a set number of events occur and some participants experience the event after the end of the study. Finally, random censoring occurs when there is some other competing event that is unrelated to the event of interest occurs which prevents participants from continuing in the study. Such events include patient withdrawal, accidental death, or patient migration. In medical research, we typically observe a mixture of type I and random censoring [213]. Clearly, censoring is a complicated and unavoidable problem, thereby requiring specialized procedures to correctly analyze time-to-event data.

Time-to-event data can be described by two frequently used and related probabilities: survival and hazard [212,213]. The survival probability at a given time is obtained by the survival function, which is given by the following formula:

$$S(t) = P(T > t)\,. \tag{11}$$

Essentially, $S(t)$ is the probability (P) that the survival time T is greater than a given time t. A standard survival function is between zero and one ($0 \leq S(t) \leq 1, for\ t \geq 0$). Specifically, the survival probability at the time origin ($t = 0$) is one and, as t increases, the survival function approaches zero. Moreover, the survival function is a non-increasing function of time t. This means that as t increases, $S(t)$ must either decrease or remain constant. Generally speaking this can be instinctive, particularly when considering death by any cause as the event of interest since everyone will experience the event at some point in time.

Another important function to describe the distribution of time-to-event data is the hazard function, $h(t)$, given by the following:

$$h(t) = \lim_{\Delta t \downarrow 0} \frac{P(T < t + \Delta t | T \geq t)}{\Delta t}\ . \tag{12}$$

Basically, $h(t)$ gives the instantaneous rate at which an event occurs at a given time t among subjects that have yet to experience the event [212,213]. As mentioned previously, the survival and hazard functions are one-to-one functions of one another. Specifically, the relationship between these two functions can be written as

$$S(t) = exp\left[-\int_0^t h(x)dx\right]. \tag{13}$$

Therefore, it is only necessary to consider one of these functions to model the survival time T.

The survival and hazard functions can also be written to condition on covariates [212,213]. The survival function conditioning on covariate $X = x$ is written as

$$S(t|x) = P(T > t|x) \ , \tag{14}$$

where $S(t|x)$ is the probability that the survival time T is greater than time t given covariate x. Essentially, this function models the survival probability stratifying the participants according to their status of x. The hazard function conditioning on a covariate is written as

$$h(t|x) = \lim_{\Delta t \downarrow 0} \frac{P(T < t + \Delta t | T \geq t, x)}{\Delta t} \ , \tag{15}$$

where $h(t|x)$ gives the instantaneous rate at which an event occurs at a given time t among subjects that have yet to experience the event given time-fixed covariate $X = x$. In genetic association analysis, the covariates are the specific genotypes of the polymorphisms and are included in the function as described in (2), (3), (4), and (6) depending on the genetic model selected for the tests. As mentioned in Section 1.3.2, when considering covariates in association tests, regression models are desirable to apply. This is no different in survival analysis. For example, one regression model commonly used in survival analysis for modelling the hazard function is the proportional hazards regression model

$$h(t|x) = h_0(t)e^{\beta x} \ , \tag{16}$$

where x is a given time-fixed covariate value, $h_0(t)$ is the baseline hazard function (the hazard function for an individual when x = 0), and e^{β} is the hazard ratio for a unit increase in the scale of x [212,213]. An important assumption of the proportional hazards regression model is the assumption of proportionality: the hazard functions are proportional overtime, or the hazard ratios are constant over time, for two subjects with a given time-fixed covariate.

The overarching aim of survival analysis is to estimate the survival or hazard function [212,213]. These functions can be estimated using several standard methods, each with their own strengths and limitations. However, there are circumstances where the standard survival analysis approaches are not appropriate to apply [214-216]. As such, I will first discuss some frequently used methods for the analysis of survival data, Kaplan-Meier estimation and Cox proportional hazards regression model, and then discuss a more advanced statistical method, the mixture cure model, which can give more informative results than the other methods under certain conditions.

1.4.2.1 Kaplan-Meier estimation

When conducting univariable analysis without considering covariates for a given subgroup, the most popular method to estimate the survival function is the Kaplan-Meier estimate, also known as the Product-Limit estimate [217]. This is a non-parametric estimation method and the Kaplan-Meier estimate of $S(t)$ is

$$\hat{S}(t) = \prod_{r:t_r<t} \frac{n_r - g_r}{n_r} \ , \tag{17}$$

where $t_1 < t_2 < \cdots < t_R$ are R ordered distinct event times in a random sample of size n, n_r is the number of participants who have not yet had the event or been censored at time t_r and g_r is the number of events at time t_r $(1 \leq r \leq R)$ [217]. From the equation, we can see that the Kaplan-Meier estimate of the survival function starts at one at the time origin and decreases as events occur [212,213]. In Chapter 3 of this thesis, specifically Section 3.5.3.1, we performed Kaplan-Meier estimation using the $survival$ package [218] in R [172].

One main feature of the Kaplan-Meier estimation method is that survival probability estimates at given time-points can be plotted against time to produce survival curves [212,213]. These curves are the easiest means to visualize survival trends. Importantly, several survival curves can be plotted simultaneously on one graph, so a visual comparison of the survival patterns for different groups is possible. While this can be useful, it is more informative to determine the statistical significance of the potential differences between survival functions [212,213]. This can be done by performing a log-rank test: a non-parametric test that determines if the distribution of the survival times is the same between two or more groups. As will be disclosed in Section 3.5.3.1, we performed a log-rank test in the project described in Chapter 3. To do this, we used the $survdiff$ function in the $survival$ package [218] in R [172]. From the output given by this function, we obtained a p-value for the corresponding test.

The Kaplan-Meier estimation is a widely-used tool for analyzing and visualizing survival pattern. Additionally, with the log-rank test, statistical inferences can be made regarding the differences in survival patterns between two or more groups. These methods are attractive because they require few assumptions and, thus, are quite flexible. However, since they are non-parametric methods, they do not provide a measure of effect [212,213]. Furthermore, these methods are not suitable for multivariable analysis as they cannot adjust for any covariates. As a result, additional survival analysis tools are required for such types of analysis.

1.4.2.2 Cox proportional hazards regression method

The Cox proportional hazards regression method is the most widely-used method for performing multivariable survival analysis [212,213,219]. It is based on the proportional hazards regression model given in (16). This is a semi-parametric method, meaning it has parametric and nonparametric components [213]. The baseline function, $h_0(t)$, is non-parametrically estimated while we have a parametric form for the coefficient of the covariate X. The model given in (16) is a univariable model, conditioning on only one covariate, X.

The multivariable proportional hazards regression model is written as

$$h(t|x) = h_0(t)e^{\beta x+\gamma\prime z} \tag{18}$$

where $h_0(t)$ is the baseline hazard function (i.e. all covariates set to 0), $Z = (Z_1, Z_2, \dots, Z_K)'$ is a vector of K covariates other than X, $\gamma' = (\gamma_1, \gamma_2, \dots, \gamma_K)$ is a vector of coefficients for the covariates in Z, and $\gamma' z = \gamma_1 z_1 + \gamma_2 z_2 + \cdots \gamma_K z_K$ denotes a linear combination of covariate values $Z = z$ and their coefficients γ. For some k between 1 and K, the hazard ratio for a unit increase in the scale of Z_k is e^{γ_k} while holding the other covariates constant. If the hazard ratio is above one, the comparison group is at a greater relative risk of experiencing the event than the reference group. Conversely, if the hazard ratio is between zero and one, the comparison group is at a reduced relative risk of experiencing the event compared to the reference group [213,219]. If the hazard ratio is one, there is no association between the survival time T and covariate Z_k.

Since the Cox proportional hazards regression model includes the unspecified baseline hazard function $h_0(t)$, in order to estimate this model, we must use a partial likelihood function [219]. This function provides the maximum partial likelihood estimators for the unknown parameters and the standard errors for these estimators from which we can calculate approximate confidence intervals and Wald-type test statistic for hypothesis testing. The Cox proportional hazards regression model does assume proportionality as described in Section 1.4.2. This assumption, in fact, can be tested by using a score test once the Cox model is fitted [220]. If this test of proportionality fails, the inference made from this model is not accurate [212,213].

Cox proportional hazards regression was applied in the study presented in Chapter 3 this thesis using R software [172]. Specifically, we used the $coxph$ function in the $survival$ package [218]. This function calculated the maximum partial likelihood estimate(s), standard error(s) of these estimates, confidence interval(s), Wald test statistic(s), and p-value(s) for each SNP and covariates included in the model. In addition, we used the $cox.zph$ function in the $survival$ package to test the proportionality assumption. This method and its specific application in this thesis will be further discussed in Section 3.5.3.1 and Section 3.5.3.4.

Both the Kaplan-Meier estimation and the Cox proportional hazards regression models are frequently used in the analysis of survival data. However, there are scenarios in which these models are not suitable to answer the specific research question. Specifically, if a disease has several important covariates that should be included in a prognostic model, Kaplan-Meier estimation cannot be applied. As for the Cox proportional hazards regression model, the proportionality assumption may be too stringent given the data. Recall that this assumption requires the hazard functions stratifying patients according to different levels of a given variable to remain proportional over time. However, the short- and long-term effects of a given covariate can be different. In fact, these circumstances can result in the survival curves stratifying patients based on such a covariate to cross each other at some time-point within the follow-up duration. The Cox proportional hazards regression model cannot adequately model the relationship between this type of covariate and the time-to-event of interest since the proportional hazards assumption is violated [215]. Consequently, in cases where several variables with possible differences in short- and long-term effects must be analyzed, cure models have been suggested as suitable methods [214,216].

1.4.2.3 Mixture cure model

As advancements in medical research and treatment continue, it is expected that a higher proportion of the population will be cured of disease [215]. As a result, Kaplan-Meier survival probability estimates will plateau at a non-zero probability at some point in time [221-223]. Specifically, the survival patterns may show an aggregation of events at the beginning of the follow-up time, but fewer events occur as time progresses until there are no more events and patients are considered statistically cured [214]. These patient populations, thus, consist of a mixture of patients who are susceptible to experiencing the event of interest and non-susceptible (i.e. cured) patients [214,215,222,224,225]. Such populations can be investigated using cure models, which estimate whether the patient is susceptible to experiencing the event (i.e. the risk of experiencing the event) and the survival probability of the patient given that the patient is susceptible [214,215,221-223,226,227]. In addition, these models can adjust for covariates.

There are several approaches to construct a cure model to estimate the survival function [214,216,225,227-237]. A widely used modelling approach [214,216,222,227,230,232-236,238], and the one used in this thesis, can be written as follows

$$S(t|x) = p(x) + \big(1 - p(x)\big)S_0(t|x)\,, \tag{19}$$

where $S(t|x) = P(T > t|x)$ is the survival probability to time t given the covariate x, $p(x) = P(T = \infty|x)$ is the probability of being cured and $S_0(t|x) = P(T > t|T < \infty, x)$ is the survival function of the time-to-event in patients who are susceptible to the event given the covariate x. The probability of being cured can be modelled using a logistic regression model where

$$p(x) = \frac{exp(a + bx)}{1 + exp(a + bx)}\,, \tag{20}$$

and e^b is the odds ratio associated with a unit increase in the scale of x for the susceptible and non-susceptible patient groups. The conditional survival function for the time-to-event can be modelled using a proportional hazards regression model

$$S_0(t|x) = exp\left[-\int_0^t h_0(u|x)du\right], \tag{21}$$

where $h_0(t|x)$ can be modelled as described above in reference to the Cox proportional hazards regression model in (15) and (17). Namely, $h_0(t|x) = h_{00}(t)e^{\beta x}$ where $h_{00}(t)$ is the baseline hazard function for the susceptible group and e^{β} is the hazard ratio for a unit increase in the scale of x in the susceptible group.

There are parametric and semi-parametric methods for estimating $p(x)$ and $S_0(t|x)$ [216,230,232]. In this thesis, we assumed a logistic regression model to model $p(x)$ and a Weibull regression model, which is a proportional hazards regression model, to model $S_0(t|x)$ [216,222]. To estimate this mixture cure model, we applied maximum likelihood estimation [222]. The likelihood function L for the observed data $\{(t_i, \delta_i, x_i), i = 1, \dots, n)\}$ of sample size n is

$$L = \prod_{i=1}^{n} f(t_i|x_i)^{\delta_i} S(t_i|x_i)^{1-\delta_i} \ . \tag{22}$$

where $f(t|x) = -\frac{\delta S(t|x)}{\delta t}$, $S(t|x)$ is given in (19), t is the observed time, δ is the censoring indicator, and x is the covariate. The parameters a, b, β, and the parameters in $h_{00}(t)$ are estimated by maximizing the likelihood function L in (22). Consequently, two risk estimates are obtained from the mixture cure model separately but simultaneously: odds ratio for the probability of being cured (e^b) and hazard ratio for the time-to-event in susceptible patients (e^β). In this case, the odds ratio compares the odds of being cured in the comparison group to the reference group considering all individuals and the hazard ratio compares the hazard functions between the comparison and reference groups among the patients who are susceptible to experiencing the event. The tests for no association for the two hypotheses were tested by the Wald test, which calculated test statistics for both the logistic regression model and the proportional hazards regression model. Specifically, we tested if $b = 0$ (or $e^b = 1$), which would indicate there is no association between the covariate x and the probability of being cured, and we tested if $\beta = 0$ (or $e^\beta = 1$), which would indicate there is no association between the covariate x and the time-to-event in the susceptible patients.

While the mixture cure model is an attractive and more informative model, it does have assumptions that need to be met for the estimates to be valid. First, the mixture cure model must be applied to a patient cohort that has a mixture of susceptible and cured patients [216]. This can be checked using Kaplan-Meier survival curves: curves that level off to a stable plateau in the long-term can be considered empirical evidence of the existence of a large proportion of cured patients. Without empirical evidence of a mixture of patients, the model reduces to a proportional hazards model. In addition, cure models benefit from long-term follow-up of the patient cohort and a small number of early censoring [214]. When these requirements are met, the mixture cure model can provide detailed and informative results regarding the risk and timing of the event of interest should the cohort be a mixture of susceptible and non-susceptible individuals.

As will be discussed in Section 3.5.3.1 and Section 3.5.3.4, to analyze the genotype data in the project described in Chapter 3 using the mixture cure model, we used R software [172]. We wrote an original code to complete this analysis which used a general purpose optimization function nlm to maximize the likelihood function. From the output of this function, we obtained the maximum likelihood estimates of the unknown parameters in the model (19) and their standard errors. From these estimates, we calculated the odds ratio and hazard ratio estimates. Additionally, since the maximum likelihood estimates are asymptotically normally distributed, we were able to calculate approximate confidence intervals and p-values.

This thesis contains two genetic association studies: one has a binary outcome variable (Chapter 2) and one has a time-to-event outcome variable (Chapter 3). The section above describes the appropriate methods of analysis for each type of outcome variable. As discussed in Section 1.3.2.3, both studies investigated all four genetic models to identify specific genotypes of the common SNPs that are associated with the given disease outcomes. This meticulousness can be quite informative and assist in increasing the confidence of the results. However, this adds a level of complexity to the research since, generally, we obtain multiple results for the same genetic variant. To isolate the most plausible genetic model, once the statistical models are fitted for each SNP under each genotype coding (Table 1.2), the fit of these models can be assessed to determine which genetic model is the most likely to be the true model.

1.4.3 Plausibility of genetic models

Considering we obtained four different results for many SNPs (one per genetic model), it could be challenging to decipher these results. To increase the accuracy of the results, we checked the plausibility of the genetic model to ensure the genetic model in which the association was identified was the best fitting model.

One criterion we used to assess the plausibility of the model was the Akaike Information Criterion (AIC) [239] for each SNP under each genetic model. The AIC is obtained using the maximized likelihood $\widehat{L^*}$, which is defined as $L^*(\hat{a}, \hat{b})$ from equation (9) where a and b are replaced by the maximum likelihood estimates $\hat{a}$ and $\hat{b}$. Specifically, the AIC is

$$AIC = 2m - 2\log\left(\widehat{L^*}\right) , \tag{23}$$

where m is the number of parameters in the model. For a given SNP, the lowest AIC value represented the model with the best fit. Accordingly, if the SNP had the highest association in the genetic model yielding the lowest AIC, the results were reported. If the SNP had the highest association in a model that was not the best fitting according to the AIC calculation, it was considered a false positive and did not undergo further investigation. This criterion for assessing the plausibility of the genetic model was applied in the study in Chapter 2, as described in Section 2.5.4.1.

Another method we employed for testing the plausibility of the genetic models is by conducting a likelihood ratio test [240]. This approach for goodness-of-fit testing compares the model under consideration with an expanded model. In the case of genetic models, the co-dominant model encompasses all the genetic models because it tests associations with each specific genotype separately. With this in mind, we conducted maximum likelihood ratio tests comparing the identified genetic model with the co-dominant model. First, we obtain the maximum likelihood for both the co-dominant model (L_0) and the model under consideration (L_1). From these values, the likelihood ratio Λ is

$$\Lambda = 2(\log(L_0) - \log(L_1)), \tag{24}$$

where L_0 and L_1 are obtained from (22) with the corresponding genetic models as defined by (2), (3), (4), and (6). Because Λ has asymptotically chi-square distribution with one degree of freedom, we can obtain a p-value when the sample size is sufficiently large. A significant difference in the goodness of fit ($p < 0.05$) would determine if we reject the proposed genetic model. If the recessive, dominant, and additive models were rejected, the co-dominant model is the plausible genetic model. As a result, after these tests, we have one appropriate association result for each SNP based on the most plausible genetic model and the results for the other genetic models were not reported. This approach for testing the goodness of fit of the different models was applied in the methods explained in Section 3.5.3.2.

1.5 Rationale, objectives, and outline

Colorectal cancer is a heterogeneous disease with significant impact on patients and health care systems. Despite extensive research, there is still much variability in this disease yet to be understood. This unknown variability may be partially explained by germline genetic variations. To that end, the overarching aim of this thesis project is to further understand potential genetic associations with colorectal cancer disease subtype formation and metastasis.

The first project in this thesis, which is described in Chapter 2, focuses on an aggressive histological variant of colorectal cancer: mucinous adenocarcinoma. The main objectives of this study were, as follows:

1) Identify specific genotypes of a genome-wide set of common SNPs (MAF ≥ 5%) independently associated with mucinous tumor histology in a colorectal cancer patient cohort (n=505) adjusting for significant baseline characteristics;

2) Identify gene-based sets of rare SNPs (MAF < 5%) associated with mucinous tumor histology in a colorectal cancer patient cohort (n=505) adjusting for significant baseline characteristics; and
3) Discuss a potential biological connection between genes/genetic variations and excessive mucin production in mucinous colorectal tumors based on current literature.

To our knowledge, this was the first genome-wide association study to assess genetic associations between both common and rare germline genetic variations and the mucinous tumor phenotype in colorectal cancer. These results may, consequently, provide unique insight into the development of mucinous tumors in colorectal cancer and offer a better understanding of the disease subtype etiology.

This project has been written in manuscript format for submission to a peer reviewed journal and a version of this manuscript is given in Chapter 2. Hence, I will give an introduction to this project and the research question, description of the methods used, summary of the results, and a discussion of these results. For ease of reading, the references for this manuscript will be given at the end of the chapter.

The second project, described in Chapter 3 of this thesis, investigated potential genetic associations with the risk and timing of metastasis of colorectal cancer. In particular, the main objectives of this study were:

1) Identify specific genotypes of a genome-wide set of common and low-frequency SNPs (MAF > 1%) associated with
    i. the long-term risk of metastasis and

ii. time-to-metastasis in patients susceptible to metastasis

in stage I-III colorectal cancer patients with MSI-L/MSS tumors (n=379) using appropriate statistical methods;

2) Perform multivariable analysis on significantly associated SNPs adjusting for appropriate baseline characteristics to assess the independence of the genetic associations; and
3) Discuss a possible biological link between the identified genes/genetic variations that may explain differential risk of/time-to-metastasis in colorectal cancer patients based on available literature.

To now, this is the first genome-wide association study investigating time-to-metastasis in colorectal cancer patients using an extensive, high-dimensional dataset. The results obtained in this study have potential clinical implications in formulating personalized treatment strategies for colorectal cancer patients based on their predicted time-to-metastasis.

A manuscript for this project has been written for submission to a peer reviewed journal. This manuscript is presented in this thesis in Chapter 3 and provides sections describing the motivation for the research, methods, results, and discussion. As with the first project, the references for this manuscript are given at the end of the chapter.

Chapter 2: Manuscript – "Associations of single nucleotide polymorphisms with mucinous colorectal cancer: genome-wide common variant and gene-based rare variant analyses"

A version of this manuscript has been prepared for submission for publishing in a peer reviewed journal. Supplementary data is provided in Appendix B.

2.1 Authors and affiliations

Michelle E. Penney[1], Patrick S. Parfrey[2], Sevtap Savas[1,3], Yildiz E. Yilmaz*[1,2,4]

1. Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada

2. Discipline of Medicine, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada

3. Discipline of Oncology, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada

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

2.2 Co-authorship statement

Michelle E. Penney: Aided in the development of the study design, conducted the statistical analyses, interpreted the results, and prepared the first draft of the manuscript.

Dr. Patrick S. Parfrey: Provided the patient data investigated in this study.

Dr. Sevtap Savas: Aided in the development of the study design and interpretation of the results, supervised the study, and revised the manuscript.

Dr. Yildiz Yilmaz: Aided in the development of the study design and interpretation of the results, reviewed all results and their interpretation, supervised the study, and revised the manuscript.

2.3 Abstract

Background: Colorectal cancer has significant impact on individuals and healthcare systems. Many genes have been identified to influence its pathogenesis. However, the genetic basis of mucinous tumor histology, an aggressive histological variant of colorectal cancer, is currently not well-known. This study aimed to identify common and rare genetic variations that are associated with the mucinous tumor phenotype.

Methods: Genome-wide single nucleotide polymorphism (SNP) data was investigated in a colorectal cancer patient cohort (n=505). Association analyses were performed for 729,373 common SNPs and 275,645 rare SNPs. Common SNP association analysis was performed using univariable and multivariable logistic regression under different genetic models. Rare-variant association analysis was performed using a multi-marker test.

Results: No associations reached the traditional genome-wide significance. However, promising genetic associations were identified. The identified common SNPs significantly improved the discriminatory accuracy of the model for mucinous tumor phenotype. Specifically, the area under the receiver operating characteristic curve increased from 0.703 (95% CI: 0.634-0.773) to 0.916 (95% CI: 0.873-0.960) when considering the most significant SNPs. Additionally, the rare variant analysis identified a number of genetic regions that potentially contain causal rare variants associated with mucinous tumor histology.

Conclusions: This is the first study applying both common and rare variant analyses to identify genetic associations with mucinous tumor histology using a genome-wide genotype data. Our results suggested novel associations with mucinous tumors. Once confirmed, these results will not only help us understand the biological basis of mucinous histology, but may also help develop targeted treatment options for mucinous tumors.

2.4 Introduction

Colorectal cancer is a global health problem and contributes substantially to worldwide cancer mortality [1]. In 2012, this disease was the 3rd most common cancer worldwide with higher rates occurring in developed countries [1]. In Canada, colorectal cancer is expected to cause 26,800 new cases and 9,400 deaths in 2017. Newfoundland and Labrador, in particular, have the highest age-standardized rates of incidence and mortality in the country [2].

Mucins are a family of high-molecular-weight glycoproteins that are widely expressed by epithelial tissues [3]. They have been identified in two forms: cell surface (transmembrane) and fully released (gel-forming) [3,4]. The gel-forming mucin-encoding genes are clustered at chromosome 11p15.5 [4,5]. These mucins, including MUC2, MUC5AC, MUC5B, and MUC6, constitute the major macromolecular components of mucus [4,5]. Among them, MUC2 is the most highly expressed one in the colorectum and is the predominant component of colorectal mucus [6-8]. MUC5B and MUC6 are highly expressed in the upper gastrointestinal (GI) tract, but low levels of both have been reported in the normal colon [8,9]. MUC5AC is highly expressed in the upper GI tract and is not expressed in the normal colon, however, abnormal expression is observed in colorectal cancer [10,11].

Mucinous adenocarcinoma is a distinct form of colorectal cancer with the defining characteristic of a high mucin component (more than 50% of the tumor volume). This subtype accounts for 5-15% of colorectal cancer cases. Compared to non-mucinous colorectal cancer, mucinous adenocarcinoma patients are typically younger and are often at an advanced stage at diagnosis [12-18]. Mucinous tumors are more likely to occur in the proximal colon [15,16,19,20] and tend to have an inferior response to systemic therapies [20,21].

Specific molecular distinctions are also seen in mucinous compared to non-mucinous colorectal tumors. For example, increased rates of *BRAF* mutations and CpG island methylator phenotype (CIMP) are observed in mucinous colorectal tumors [22]. In addition, overexpression of MUC2, strong ectopic expression of gastric MUC5AC, and decreased p53 expression in mucinous tumors are reported in the literature [23,24]. Mucinous and non-mucinous tumors also appear to have differences in genome-wide gene expression patterns [18]. Some of the upregulated genes in mucinous tumors are involved in cellular differentiation and mucin metabolism, which are characteristics biologically relevant to the phenotype [18]. While some differences between mucinous and non-mucinous colorectal cancers are well recognized, the prognostic importance of a high mucin component has been controversial [14-16,20,21,25-30].

Most studies investigating genetic characteristics of mucinous colorectal tumors examined single or a limited number of candidate genes [6,31,32]. Conversely, this study aimed to comprehensively identify common and rare genetic polymorphisms that may be influencing the production of mucin or formation of the mucinous tumor phenotype. To do so, we applied a genome-wide approach to identify genes and genetic regions that are associated with the risk of developing mucinous colorectal tumors.

2.5 Methods

2.5.1 Ethics statement

Patient consents were obtained by the Newfoundland Colorectal Cancer Registry (NFCCR) at the time of recruitment. If the patient was deceased, consent was sought from a close relative [33]. Ethics approval for this study was obtained from the Health Research Ethics Board (HREB; #15.043). Since this is a secondary use of data study, no patient consent specifically for this study was required.

2.5.2 Patient cohort

The study cohort was a subgroup of the NFCCR and consisted of 505 Caucasian patients. Both the NFCCR and the study cohort were described in detail in other publications [34,35]. In short, the NFCCR recruited 750 colorectal cancer patients in Newfoundland and Labrador collected between 1999 and 2003. All diagnoses were confirmed by pathological examination. Out of 750 patients, 505 patients constituted the study cohort as explained below.

2.5.3 Genotype data

The genotype data used in this study was explained in Xu *et al.* (2015) [35]. In short, DNA samples of 539 patients were subject to whole-genome single nucleotide polymorphism (SNP) genotyping using the Illumina Omni1-Quad human SNP genotyping platform. These patients were included into the genetic analysis because of the availability of their outcome and clinical data as well as the germline DNAs extracted from peripheral blood samples. The quality control analysis and filtering for this data included removing SNPs whose frequencies deviated from Hardy-Weinberg equilibrium, SNPs that had $> 5\%$ missing values, and patients with discordant sex information, accidental duplicates, divergent or non-Caucasian ancestry, and first, second, or third degree relatives [35]. In this genotype data, there were 505 patients with 729,373 common SNPs (minor allele frequency; MAF ≥ 0.05) and 275,645 rare SNPs (MAF < 0.05) that were included in this study. During this study, management and handling of these genotype data was done using PLINK v. 1.07 [36].

2.5.4 Statistical analysis

All statistical analysis was performed using R v. 3.1.3 [37]. Correction for multiple testing was not applied to the results as this is an exploratory study and we did not want to increase false negative rate due to conservative corrections. While this increases the chances of obtaining false positives, we believe replication of these results in other studies will assist in reducing the potential false positive findings.

2.5.4.1 Common SNP analysis

Univariable logistic regression analysis: Univariable logistic regression analysis was performed on each common SNP (MAF $\geq$ 5%) to determine if individual SNPs were significantly associated with the mucinous tumor phenotype. For each SNP, the additive, co-dominant, dominant, and recessive genetic models were applied. Consequently, we report the 10 SNPs with the highest level of significance in each genetic model (Supplementary Tables S1-S4).

Selection of baseline variables and multivariable logistic regression analysis: In order to select significant baseline factors to adjust for in the multivariable analyses, we first examined the variables shown in Table 2.1 using univariable logistic regression models. Factors that had a p-value less than 0.1 were then included in a forward stepwise variable selection method. In addition, although there appeared to be a non-significant association between tumor histology and grade in the univariable analysis, tumor grade was still included in the multivariable model as has been shown to be linked to tumor histology [25,38]. As a result, the baseline characteristics in the final models were sex, age at diagnosis, stage, and tumor location based on the 0.1 level of significance, and tumor grade (Supplementary Table S5). The 10 SNPs with the highest level of significance under each genetic model in the univariable logistic regression analysis were analyzed using the multivariable logistic regression model adjusting for the selected baseline characteristics (Supplementary Tables S1-S4).

Table 2.1 Baseline features of the study cohort and the results of univariable logistic regression analysis

Characteristics		Mucinous No. (%)	Non-mucinous No. (%)	OR (95% CI)	p-value
Age[a]	≤60	20 (9)	203 (91)		
	60-65	17 (18)	78 (82)	2.21 (1.09-4.44)	0.025
	>65	20 (11)	167 (89)	1.22 (0.63-2.34)	0.558
Sex	Female	29 (15)	169 (85)		
	Male	28 (9)	279 (91)	0.58 (0.34-1.02)	0.057
Location	Colon	47 (14)	287 (86)		
	Rectum	10 (6)	161 (94)	0.38 (0.18-0.74)	0.007
Stage	I	3 (3)	90 (97)		
	II	27 (14)	169 (86)	4.79 (1.64-20.45)	0.012
	III	19 (11)	147 (89)	3.88 (1.28-16.83)	0.033
	IV	8 (16)	42 (84)	5.71 (1.57-27.09)	0.013
Grade	Well/moderately diff.	48 (10)	416 (90)		
	Poorly diff.	7 (19)	30 (81)	2.02 (0.78-4.62)	0.115
	Unknown	2	2		
MSI status	MSI-low/MSS	49 (11)	382 (89)		
	MSI-high	6 (11)	47 (89)	1.00 (0.37-2.29)	0.992
	Unknown	3	18		
Lymphatic invasion	Absent	31 (10)	267 (90)		
	Present	23 (14)	144 (86)	1.38 (0.77-2.44)	0.278
	Unknown	3	37		
BRAF V600E mutation	Absent	45 (11)	366 (89)		
	Present	9 (19)	38 (81)	1.93 (0.83-4.09)	0.104
	Unknown	3	44		

[a]The age at diagnosis was separated into 3 groups: ≤60, 60-65, and >65 since this particular grouping gave the most efficient estimates with no significant change in the results when considering slightly different groupings. CI: confidence interval, diff.: differentiated, MSI: microsatellite instability, MSS: microsatellite stable, No: number, OR: odds ratio (compares the odds of having mucinous tumors with the corresponding factor level to the odds of having mucinous tumors with the reference factor level).

Plausibility of the genetic models: It is common in genetic association studies that only one genetic model is applied. In this study we applied all four genetic models and assessed the plausibility of the genetic model under which the SNP was identified. To do this, we used the Akaike Information Criterion (AIC) calculations to compare the fit of four different genetic models per SNP under the multivariable logistic regression model in (7). The genetic model with the smallest AIC estimate was considered to be the most plausible genetic model. We first ranked the SNPs based on their p-value obtained in the multivariable model with the genetic model under which the SNP was identified (Supplementary Table S6). Then, we excluded those SNPs that were not identified in their plausible genetic model. Of note, we present in this manuscript only the 10 SNPs that have the highest association significance levels under the multivariable logistic regression models that were identified in their most plausible genetic model. We refer to these SNPs as “the top 10 SNPs”.

Assessing the discriminatory accuracy of the estimated models: We aimed to check the ability of the multivariable models of the top 10 SNPs to discriminate between mucinous and non-mucinous tumors. A well-known method for testing the discriminatory accuracy of a model is using a receiver operating characteristic (ROC) curve [39-41]. Calculating the area under the curve (AUC) of the ROC curve for the given models provides a single numeric representation for the performance of the model [40,42,43]. Comparing the AUC values and their corresponding confidence intervals provides a method for determining if one model is significantly superior to another in diagnostic accuracy [41,44].

ROC curve analysis was performed by calculating the AUC using the pROC package in R [45]. AUC estimates were calculated for (i) the model conditioning only on the baseline characteristics, (ii) the model conditioning on only the top SNPs, and (iii) the model conditioning on the baseline characteristics and the top SNPs. Comparing the AUC estimates, specifically the 95% confidence intervals, between these three models can quantify the differences in the capacity of the models to distinguish mucinous and non-mucinous tumors.

2.5.4.2 Rare variant analysis

SKAT-O analysis: SKAT-O [46] test statistic was used to test the associations between the rare variants and mucinous tumor histology. For this analysis, we prioritized gene-based regions including 5 kb long sequences before and after each gene. To do so, we first obtained genome location information for genome-wide gene-based regions (for the reference genome GRCh37.p13) using the biomaRt tool [47] in the Ensembl database [48]. The SNP information within these regions were then retrieved from the patient genome-wide data and used as the region-based SNP-sets in SKAT-O. During this analysis, each SNP was assigned to one gene-based region only. As a result, when a gene is located in close proximity to another gene, the second gene-based region does not include the SNPs that are analyzed in the first gene-based region. This limits redundancy since no SNP is analyzed more than once. For this analysis, only the additive genetic model was considered as using multiple genetic models is not a practical option for SKAT-O. The associations of gene regions were examined in multivariable models adjusting for sex, age at diagnosis, stage, tumor location, and tumor grade.

2.5.5 Bioinformatics analysis

Potential regulatory consequences of the identified SNPs were examined through RegulomeDB (http://www.regulomedb.org/) [49]. Ensembl [48] database was used to retrieve information related to the genes identified in the common and rare variant analysis.

2.6 Results

The demographic and clinicopathological information for the sample population is shown in Table 2.1. We observed a non-significant association of tumor histology with age at diagnosis (> 65 versus ≤ 60), grade, microsatellite instability (MSI) status, lymphatic invasion (LI), and *BRAF* V600E mutation; a moderately significant association with stage, sex, and age at diagnosis between 60-65 versus ≤ 60; and a strongly significant association with tumor location (Table 2.1). In this cohort, there was a trend for female sex having increased risk of mucinous tumors. As expected, the proportion of mucinous tumors was higher in colon cancer patients compared to rectum cancer patients and in stage II-IV patients compared to stage I patients (Table 2.1).

2.6.1 Common SNP analysis

None of the associations in this analysis reached the traditional genome-wide significance level ($p < 5 \times 10^{-8}$), but each genetic model identified promising associations.

After the univariable analysis, there were 33 SNPs that were nominally associated with the mucinous tumor phenotype (Supplementary Tables S1-S4**).** Associations of two SNPs (rs11216624 & rs17712784) were identified in both the dominant and co-dominant genetic models; one SNP (rs7314811) was detected in the additive, recessive, and co-dominant genetic models; and three SNPs (rs4843335, rs10511330, & rs16822593) were detected in both the additive and dominant genetic models. The estimates obtained in the univariable analysis did not change significantly when the models were adjusted for the baseline characteristics (Supplementary Tables S1-S4).

As explained in the Methods section, the Akaike Information Criterion (AIC) estimates (Supplementary Table S6) were used to determine the most plausible genetic models for each of 33 SNPs. The ten SNPs with the smallest p-value in the multivariable analysis under the most plausible genetic models were further prioritized (i.e., the top 10 SNPs). The results of the univariable and the multivariable logistic regression analyses for these top 10 SNPs are summarized in Table 2.2. Seven of these SNPs were located within gene sequences. These genes were quite diverse and belong to a variety of biological processes and pathways (Table 2.3).

Before the ROC analysis, the linkage disequilibrium (LD) among the top 10 SNPs was assessed using patient genotype data. These calculations indicated that rs13019215 and rs12471607 were in complete pairwise LD (r^2 = 1). The SNPs rs4837345 and kgp10457679 were also in high LD with each other, as well as rs10511330 and rs16822593 ($0.99 \leq r^2 \leq 1.0$). Therefore, we arbitrarily selected one SNP per SNP set in

Table 2.2 Top ten promising SNPs identified in univariable analysis and the subsequent multivariable analysis under their plausible genetic models

					Univariable		**Multivariable**	
Genomic Location	**SNP ID (Genotype[a])**	**Gene[b]**	**Information in RegulomeDB**	**Plausible Model[c]**	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
Chr6:110750552	rs9481067 (GG)	*SLC22A16*	ND	Recessive	4.17 (2.33-7.43)	1.24E-06	4.75 (2.53-8.95)	1.24E-06
Chr3:114121019	rs10511330 (CT + CC)	*ZBTB20*	Minimal binding evidence	Dominant	3.77 (2.06-6.81)	1.24E-05	4.85 (2.54-9.23)	1.40E-06
Chr3:114117327	rs16822593 (AG + AA)	*ZBTB20*	ND	Dominant	3.70 (2.02-6.68)	1.59E-05	4.83 (2.53-9.20)	1.50E-06
Chr2:179860562	rs13019215 (TC + TT)	*CCDC141*	ND	Dominant	0.27 (0.14-0.48)	1.56E-05	0.23 (0.12-0.43)	8.20E-06
Chr2:179867985	rs12471607 (TC + TT)	*CCDC141*	ND	Dominant	0.27 (0.14-0.48)	1.65E-05	0.23 (0.12-0.43)	8.42E-06
Chr5:80483574	rs716897 (CT + CC)	*RASGRF2*	Minimal binding evidence	Dominant	0.27 (0.15-0.47)	5.33E-06	0.26 (0.14-0.47)	1.12E-05
Chr16:86077637	rs4843335 (AG + AA)	intergenic	Minimal binding evidence	Dominant	4.11 (2.11-7.79)	2.06E-05	4.67 (2.98-9.34)	1.48E-05
Chr6:118634698	rs11968293 (CA + CC)	*SLC35F1*	Minimal binding evidence	Dominant	0.28 (0.16-0.50)	1.27E-05	0.26 (0.14-0.48)	1.48E-05
Chr9:131923949	rs4837345 (TT)	intergenic	Minimal binding evidence	Recessive	4.72 (2.40-9.05)	4.00E-06	4.56 (2.24-9.11)	1.97E-05
Chr9:131930494	kgp10457679/ rs10819474[d] (CC)	intergenic	Likely to affect binding and linked to expression of a gene target	Recessive	4.72 (2.40-9.05)	4.00E-06	4.56 (2.24-9.11)	1.97E-05

[a]Risk increasing/decreasing genotype. [b]Based on Ensembl [48] or dbSNP databases [50]. [c]Under the recessive genetic model, minor allele homozygous patients are compared to major allele homozygous and heterozygous patients combined. Under the dominant genetic model, minor allele homozygous and heterozygous patients are combined and compared to major allele homozygous patients. [d]The rs number for the kgp10457679 polymorphism was obtained from the UCSC genome browser [51]. *Multivariable models adjusted for sex, age at diagnosis, stage, tumor location, and tumor grade. Patients with missing/unknown data for any of these variables were excluded from the analysis. Chr: chromosome, CI: confidence interval, ND: data not available at RegulomeDB, OR: odds ratio (compares the odds of having mucinous tumors with the specified genotype(s)[a] to the odds of having mucinous tumors with the reference (other) genotype(s)).

Table 2.3 Genes identified in the common and rare analyses

Gene Symbol[a]	Gene Name[b]	Function
SLC22A16	solute carrier family 22 member 16	codes for a human ʟ-carnitine transporter protein hCT2. hCT2 has been shown to have undetectable expression in a colon cancer cell line. [52,53]
CCDC141	coiled-coil domain containing 141	codes for a protein that plays a critical role in centrosome positioning and movement, particularly radial migration. Centrosome aberrations have been shown to be present in early-stage colorectal cancers and could contribute to chromosomal instability. [54,55]
SLC35F1	solute carrier family 35 member F1	codes for a member of the solute carrier family 35, a family of nucleotide sugar transporters. [56]
ZBTB20	zinc finger and BTB domain containing 20	codes for a transcriptional repressor. Upregulation of ZBTB20 has been shown to promote cell proliferation in non-small cell lung cancer and is a potential druggable target for the disease. Similarly, overexpression of ZBTB20 has been associated with poor prognosis in patients with hepatocellular carcinoma. [57-59]
RASGRF2	Ras protein specific guanine nucleotide releasing factor 2	codes for a signalling molecule. RasGRF2 contains regulatory domains for both Ras and Rho GTPases, suggesting it can influence both pathways. The Rho pathway has been thought to be involved in cell migration, while the Ras pathway has been thought to be involved in cell proliferation and survival, which are all processes related to cancer. [60,61]
SEC24B	SEC24 homolog B, COPII coat complex component	codes for a protein that is a part of the COPII vesicle coat, facilitating molecular transport from the endoplasmic reticulum to the Golgi apparatus. It has been suggested that alterations in vesicle trafficking proteins may be facilitators of epithelial carcinogenesis.[62,63]
CCDC109b	coiled-coil domain containing 109B	also known as *MCUb*. This gene codes a protein that interacts with the mitochondrial calcium transporter protein, CCDC109a/MCU, reducing the activity of the transporter. Calcium homeostasis in mitochondria may regulate cell death pathways.[64,65]
LINC00596	long intergenic non-protein coding RNA 596	no literature data available.
SEC24B-AS1	SEC24B antisense RNA 1	long non-coding RNA (lncRNA) that is involved in gene expression regulation. [66]
RP11-564A8.8	NA	no literature data available.
FAM87A	family with sequence similarity 87 member A	no literature data available.

[a]According to Ensembl database [48]. [b]According to HUGO Gene Nomenclature Committee (HGNC) [67]. NA = Not available.

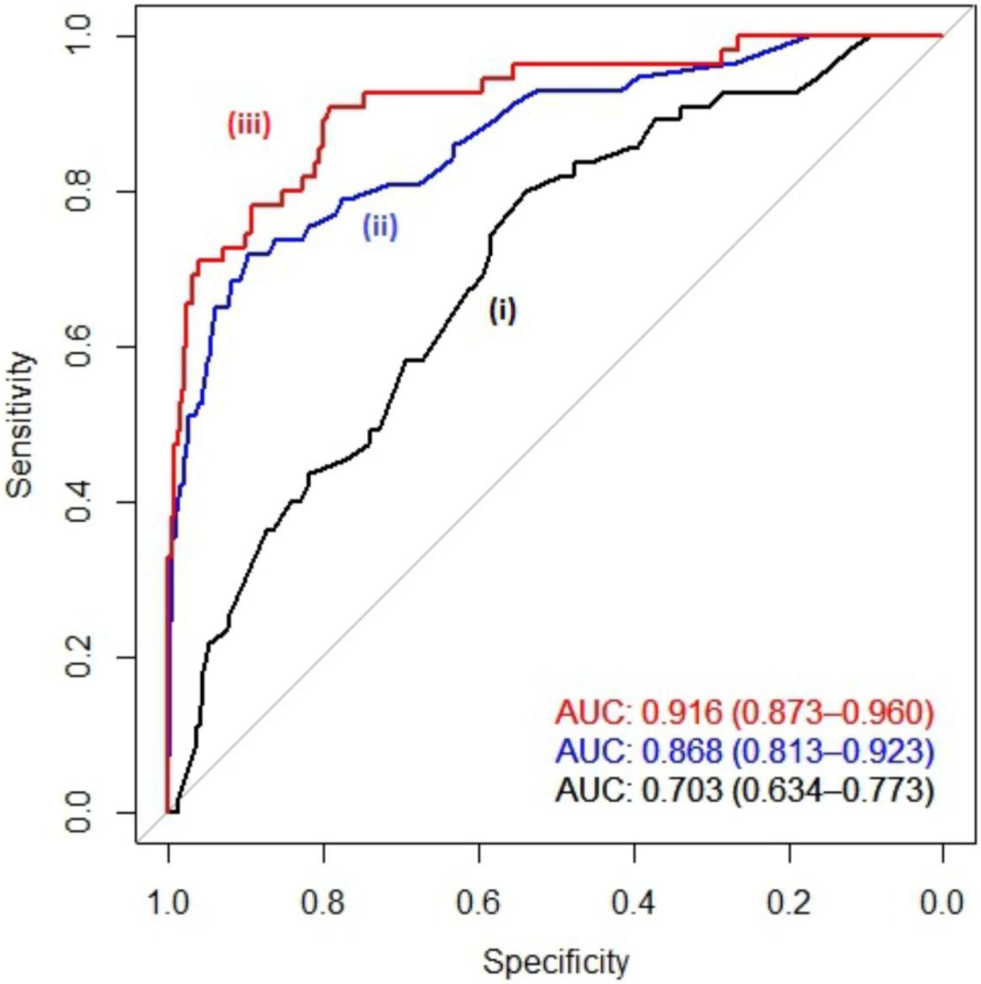
high LD, which left the following SNPs for the ROC analysis: rs9481067, rs10511330, rs13019215, rs716897, rs4843335, rs11968293, and kgp10457679.

Figure 2.1 shows the ROC curves comparing the accuracy of the models to discriminate between mucinous and non-mucinous tumors. The model (iii) including both the baseline characteristics and the SNPs (area under the ROC curve (AUC) = 0.916, CI: 0.873-0.960) had the most discriminatory accuracy followed by model (ii) including only the SNPs (AUC = 0.868, CI: 0.813-0.923) and model (i) including only the baseline characteristics (AUC = 0.703, 95% CI: 0.634-0.773). Since the confidence intervals of models (i) and (iii) do not overlap, we can confidently claim that there is a statistically significant improvement in the discriminating accuracy of the model containing the SNPs [41,44]. This also suggests that these SNPs explain some of the variation between mucinous and non-mucinous tumor histology.

2.6.2 Rare SNP Analysis

In the gene region-based rare variant analysis, we investigated 29,966 regions in the patient cohort using the multivariable SKAT-O method. Table 2.3 and Table 2.4 summarize the most significant regions ($p < 10^{-4}$) that potentially contain causal rare variants associated with the mucinous tumor phenotype. The number of variants aggregated in these gene-based regions varied from 5 - 10. While three of these regions (including *SEC24B*, *SEC24B-AS1*, and *CCDC109B*) were located close to each other on chromosome 4, other regions come from different parts of the genome (Table 2.4).

Figure 2.1 ROC curves and corresponding AUC values for multivariable models.



Due to high LD among some of the top 10 SNPs, ROC analysis was performed on only the following SNPs: rs9481067, rs10511330, rs13019215, rs716897, rs4843335, rs11968293, and kgp10457679.

AUC: area under the ROC curve, CI: confidence interval, LD: linkage disequilibrium, ROC: receiver operator characteristic.

Table 2.4 Most significant gene regions identified from SKAT-O multivariable analysis.

Genomic Location[a]	**Gene**[b]	**Description**[c]	**Other genes in the gene-based region**[d]	**# of SNPs**	**SNPs**	**p-value**
Chr4:110349928-110467052	*SEC24B*	protein coding	*SEC24B-AS1* (partial sequence)	5	rs10516557, kgp21293502, rs10003981, rs17040515, rs17040519	1.81E-05
Chr4:110476361-110614874	*CCDC109B*	protein coding	*CDC42P4* (pseudogene: partial sequence), *HIGD1AP14* (pseudogene; full length), *CASP6* (partial sequence)	6	rs17619262, rs7654187, rs6831048, rs17619310, rs9997940, rs1053680	3.29E-05
Chr14:24386456-24408777	*LINC00596*	long intergenic non-protein coding RNA	*DHRS4-AS1* (partial sequence)	6	rs8010486, rs1159372, rs10135026, rs8005541, rs8019962, kgp19564619	3.34E-05
Chr4:110263631-110359973	*SEC24B-AS1*	noncoding RNA; antisense RNA	*RBMXP4* (pseudogene; full length), *SEC24B* (partial sequence)	7	rs10031399, rs17040364, rs17040369, rs11098033, rs17040401, rs12648138, rs11098035	4.21E-05
Chr1:207074273-207084738	*RP11-564A8.8*	pseudogene	*IL24* (partial sequence), *FAIM3* (partial sequence), *FCMR* (partial sequence)	10	rs3093428, kgp15249933, kgp15191074, rs3093447, kgp22852559, rs3093434, rs3093437, rs3093438, rs3093440, rs41304091	5.47E-05
Chr8:320931-338174	*FAM87A*	non-coding RNA	-	7	rs4527844, kgp20525414, kgp20198205, rs11785854, rs7461388, rs17064450, rs17064458	6.58E-05

[a]These genomic locations describe the region containing the gene as well as 5 kb long sequences before and after the gene. [b]Based on the information in the UCSC database [51]. [c]NCBI's Gene Entrez database [66]. [d]In some cases, the gene regions examined also contained sequences of other genes. Chr = chromosome.

2.7 Discussion

Mucinous tumors are considered an aggressive type of colorectal tumors that are poorly understood [17,19,68]. While their role in prognosis is not well established, several studies suggested these tumors are associated with poorer prognosis when compared to non-mucinous tumors [20,21,27,28,30]. Identification of genes and genetic variations that can have a role in mucinous tumor development, therefore, has both scientific (e.g. dissecting the biology behind the mucinous tumor histology) as well as clinical value (e.g. biological information gained may assist with development of targeted treatment for this cancer subtype). Accordingly, for the first time with this study, we examined associations of both common and rare variants with the risk of developing mucinous tumors using a genome-wide dataset.

While our results did not reach the conservative genome-wide significance level, promising associations were detected in both the common and rare variant analyses. In common SNP analysis, we identified seven unlinked polymorphisms that significantly increased our capacity to discriminate between mucinous and non-mucinous tumors (Figure 2.1, Table 2.2). Their effects on mucinous histology were independent from the effects of the baseline variables (Figure 2.1, Table 2.2). It is possible these polymorphisms (or others in high LD with them [Supplementary Table S7], including three additional SNPs shown in Table 2.2) are biologically linked to the mucinous tumor phenotype or mucin production. Since there was no reported functional consequence of these SNPs in the literature, we searched the RegulomeDB database [49] for their potential regulatory characteristics. As of October 2017, the only SNP with a predicted/reported regulatory function in this database was kgp10457679 (rs10819474) (RegulomeDB score = 1f). This intergenic SNP is categorized as an expression quantitative trait locus (eQTL)/transcription factor (TF) binding/DNAse peak site, with a likely role of influencing the expression of target genes (Supplementary Table S8). Specifically, *PPP2R4* is noted as the eQTL for this SNP. PPP2R4 is a tumor suppressor protein [69] which has been shown to have low activity in a large portion of a small cohort of colorectal tumors [70] and is associated with shorter survival times in metastatic colorectal cancer patients [71]. A potential link between PPP2R4 and mucinous tumor histology should be examined in further studies. Overall, all the novel loci identified by the common variant analysis are interesting candidates in examination of mucinous tumor histology.

Typical association studies, such as the common variant analysis, focus on a variant-by-variant approach, which can be underpowered for rare variants. It has been suggested that gene/region-based approaches can be useful in increasing the power under these circumstances where the direct effects of multiple variants on a phenotype can instead be examined [72]. In this study, we performed the first rare variant analysis to explore gene regions that may have a role in mucinous tumor formation using SKAT-O [46]. SKAT-O is a multi-marker association test which has reasonable type I error rate and is a powerful test under many scenarios [46]. In our study, this method identified a number of gene-based regions that may harbor rare variants associated with mucinous tumors (Table 2.3, Table 2.4). Interestingly, three of the gene-based regions in Table 2.4 (*SEC24B*, *SEC24B-AS1*, and *CCDC109B*-based regions) were located in a 341,243 bp long genomic region on chromosome 4q. Since we assigned each SNP to only one gene region, these results suggest that these three gene regions are associated with mucinous colorectal tumors independent of each other. A search on the RegulomeDB database [49] indicated that one of the SNPs in *LINC00596* (rs8005541) could have a strong regulatory function (RegulomeDB score = 1f). This variant is located in an eQTL and seems to affect the expression of two nearby genes; *DHRS4* and *DHRS4L2*. These two genes are a part of a gene cluster on chromosome 14 that code for dehydrogenases/reductases [73] and have not been previously linked to mucinous colorectal tumors. Similarly, none of the genes in Table 2.4 had a previously identified connection to the risk of developing mucinous tumors. In conclusion, these regions, genes, or SNPs, alone or in combination, may be influential in the development of mucinous tumor histology and should be explored further.

Several strengths and limitations of this study should be mentioned. Studying the mucinous tumor phenotype is inherently challenging since it is not frequently detected. Despite this and the large number of SNPs/gene-based regions investigated, this study identified promising variants and genetic regions that may have a biological connection to mucinous tumor histology. We are aware that our results need to be replicated in independent cohorts and remain to be verified. Of note, SNPs and genetic regions we report are different than the MUC genes, which are the typical candidate genes for mucin production and mucinous histology. In the common variant analysis, the recessive and co-dominant models yielded some high odds ratio estimates but also wide confidence intervals (as expected, as these are the models with relatively low power). Consequently, the interpretation of these results should be made with caution. Additionally, SKAT-O is a robust test and an attractive choice for rare variant analysis, however, it cannot determine which SNPs or how many SNPs within a SNP-set are truly associated with the phenotype. Also, in the rare variant analysis, due to the assignment of one SNP to one gene region, there could be some genes whose associations may have been missed. Finally, in contrast to previous studies, we used a comprehensive genome-wide SNP genotype data. However, analysis of a more comprehensive data (such as those obtained by whole genome sequencing) would be desirable. This is particularly true for rare variants as most genotyping technologies target primarily common SNPs. Future studies should focus their efforts on sequence data to obtain a more complete dataset of rare as well as common variants.

In conclusion in this study, we performed the first genome-wide association study investigating common and rare SNPs in colorectal cancer patients to identify novel genetic associations with mucinous tumor histology. We identified novel, promising, and independent associations of specific common SNP genotypes with the risk of developing mucinous tumors. Furthermore, these SNPs significantly improved the discriminatory accuracy of the multivariable model to distinguish between mucinous and non-mucinous tumors. In addition, we detected novel promising associations between gene-based sets of rare SNPs and mucinous tumors. The results of this study, once replicated in other cohorts, can contribute further information to the molecular characteristics of this under-studied but clinically important colorectal cancer histological variant.

2.8 Acknowledgements

We thank the patients and families that participated in NFCCR and all the NFCCR personnel/investigators who contributed to NFCCR. This work was primarily funded by the Ignite fund - Research and Development Corporation (RDC) Newfoundland and Labrador (NL) awarded to Y.E. Yilmaz. Other funds that supported NFCCR and helped collect the data used in this study are gratefully acknowledged: Medical Research Fund-Cox Award-2010; Research and Development Corporation of Newfoundland (RDC; contract number: 5404.1201.102), Canadian Institute of Health Research (CIHR; RPP-operating fund; FRN: 110045), CIHR fund for the Colorectal Cancer Interdisciplinary Health Research Team at the University of Toronto and Memorial University, the National Cancer Institute of Canada (grants 18223 and 18226) and the Atlantic Innovation Fund for the Interdisciplinary Research Team in Human Genetics. M. Penney is supported by a Translational and Personalized Medicine Initiative (TPMI) fellowship. S. Savas is a senior scientist at Beatrice Hunter Cancer Research Institute (BHCRI). The funding sources were not involved in the study design; in the collection, analysis or interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

2.9 References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359-E386.

2. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian cancer statistics 2017. *Toronto, ON: Canadian Cancer Society*. 2017 Available at: cancer.ca/Canadian-Cancer-Statistics-2017-EN.pdf.

3. Moniaux N, Escande F, Porchet N, Aubert JP, Batra SK. Structural organization and classification of the human mucin genes. *Front Biosci*. 2001;6:d1192-1206.

4. Desseyn J, Aubert J, Porchet N, Laine A. Evolution of the large secreted gel-forming mucins. *Mol Biol and Evol*. 2000;17(8):1175-1184.

5. Desseyn J, Buisine M, Porchet N, Aubert J, Degand P, Laine A. Evolutionary history of the 11p15 human mucin gene family. *J Mol Evol*. 1998;46(1):102-106.

6. Okudaira K, Kakar S, Cun L, et al. MUC2 gene promoter methylation in mucinous and non-mucinous colorectal cancer tissues. *Int J Oncol*. 2010;36(4):765-775.

7. Johansson MEV, Larsson JMH, Hansson GC. The two mucus layers of colon are organized by the MUC2 mucin, whereas the outer layer is a legislator of host-microbial interactions. *Proc Natl Acad Sci U S A*. 2010;108:4659-4665.

8. Ho SB, Niehans GA, Lyftogt C, et al. Heterogeneity of mucin gene expression in normal and neoplastic tissues. *Cancer Res*. 1993;53(3):641-651.

9. Toribara NW, Roberton AM, Ho SB, et al. Human gastric mucin. identification of a unique species by expression cloning. *J Biol Chem*. 1993;268(8):5879-5885.

10. Biemer-Hüttmann A, Walsh MD, McGuckin MA, et al. Immunohistochemical staining patterns of MUC1, MUC2, MUC4, and MUC5AC mucins in hyperplastic polyps, serrated adenomas, and traditional adenomas of the colorectum. *J Histochem Cytochem*. 1999;47(8):1039-1048.

11. Bartman AE, Serson SJ, Ewing SL, et al. Aberrant expression of MUC5AC and MUC6 gastric mucin genes in colorectal polyps. *Int J Cancer*. 1999;80(2):210-218.

12. Wu C, Tung S, Chen P, Kuo Y. Clinicopathological study of colorectal mucinous carcinoma in taiwan: A multivariate analysis. *J Gastroenterol Hepatol*. 1996;11(1):77-81.

13. Odone V, Chang L, Caces J, George SL, Pratt CB. The natural history of colorectal carcinoma in adolescents. *Cancer*. 1982;49(8):1716-1720.

14. Chew M, Yeo SE, Ng Z, et al. Critical analysis of mucin and signet ring cell as prognostic factors in an asian population of 2,764 sporadic colorectal cancers. *Int J Colorectal Dis*. 2010;25(10):1221-1229.

15. Papadopoulos VN, Michalopoulos A, Netta S, et al. Prognostic significance of mucinous component in colorectal carcinoma. *Tech Coloproctol*. 2004;8(1):s123-s125.

16. Kang H, O'Connell BJ, Maggard AM, Sack J, Ko YC. A 10-year outcomes evaluation of mucinous and signet-ring cell carcinoma of the colon and rectum. *Dis Colon Rectum*. 2005;48(6):1161-1168.

17. Consorti F, Lorenzotti A, Midiri G, Di Paola M. Prognostic significance of mucinous carcinoma of colon and rectum: A prospective case-control study. *J Surg Oncol*. 2000;73(2):70-74.

18. Melis M, Hernandez J, Siegel EM, et al. Gene expression profiling of colorectal mucinous adenocarcinomas. *Dis Colon Rectum*. 2010;53(6):936-943.

19. Nozoe T, Anai H, Nasu S, Sugimachi K. Clinicopathological characteristics of mucinous carcinoma of the colon and rectum. *J Surg Oncol*. 2000;75(2):103-107.

20. Catalano V, Loupakis F, Graziano F, et al. Mucinous histology predicts for poor response rate and overall survival of patients with colorectal cancer and treated with first-line oxaliplatin- and/or irinotecan-based chemotherapy. *Br J Cancer*. 2009;100(6):881-887.

21. Negri FV, Wotherspoon A, Cunningham D, Norman AR, Chong G, Ross PJ. Mucinous histology predicts for reduced fluorouracil responsiveness and survival in advanced colorectal cancer. *Ann Oncol*. 2005;16(8):1305-1310.

22. Tanaka H, Deng G, Matsuzaki K, et al. BRAF mutation, CpG island methylator phenotype and microsatellite instability occur more frequently and concordantly in mucinous than non-mucinous colorectal cancer. *Int J Cancer*. 2006;118(11):2765-2771.

23. Hanski C, Tiecke F, Hummel M, et al. Low frequency of p53 gene mutation and protein expression in mucinous colorectal carcinomas. *Cancer Lett*. 1996;103(2):163-170.

24. Park SY, Lee HS, Choe G, Chung JH, Kim WH. Clinicopathological characteristics, microsatellite instability, and expression of mucin core proteins and p53 in colorectal mucinous adenocarcinomas in relation to location. *Virchows Archiv*. 2006;449(1):40-47.

25. Farhat MH, Barada KA, Tawil AN, Itani DM, Hatoum HA, Shamseddine AI. Effect of mucin production on survival in colorectal cancer: A case-control study. *World J Gastroenterol*. 2008;14(45):6981-6985.

26. Nitsche U, Zimmermann A, Späth C, et al. Mucinous and signet-ring cell colorectal cancers differ from classical adenocarcinomas in tumor biology and prognosis. *Ann Surg*. 2013;258(5):775-783.

27. Numata M, Shiozawa M, Watanabe T, et al. The clinicopathological features of colorectal mucinous adenocarcinoma and a therapeutic strategy for the disease. *World J Surg Oncol*. 2012;10:109-109.

28. Verhulst J, Ferdinande L, Demetter P, Ceelen W. Mucinous subtype as prognostic factor in colorectal cancer: A systematic review and meta-analysis. *J Clin Pathol*. 2012;65(5):381-388.

29. Nitsche U, Friess H, Agha A, et al. Prognosis of mucinous and signet-ring cell colorectal cancer in a population-based cohort. *J Cancer Res Clin Oncol*. 2016;142(11):2357-2366.

30. Park JS, Huh JW, Park YA, et al. Prognostic comparison between mucinous and nonmucinous adenocarcinoma in colorectal cancer. *Medicine*. 2015;94(15):e658.

31. Hanski C. Is mucinous carcinoma of the colorectum a distinct genetic entity? *Br J Cancer*. 1995;72(6):1350-1356.

32. Kim DH, Kim JW, Cho JH, et al. Expression of mucin core proteins, trefoil factors, APC and p21 in subsets of colorectal polyps and cancers suggests a distinct pathway of pathogenesis of mucinous carcinoma of the colorectum. *Int J Oncol*. 2005;27:957-964.

33. Green RC, Green JS, Buehler SK, et al. Very high incidence of familial colorectal cancer in newfoundland: A comparison with ontario and 13 other population-based studies. *Fam Cancer*. 2007;6(1):53-62.

34. Woods MO, Hyde AJ, Curtis FK, et al. High frequency of hereditary colorectal cancer in newfoundland likely involves novel susceptibility genes. *Clin Cancer Res*. 2005;11(19):6853.

35. Xu W, Xu J, Shestopaloff K, et al. A genome wide association study on newfoundland colorectal cancer patients' survival outcomes. *Biomarker Res*. 2015;3(1):6.

36. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*. 2007;81(3):559-575.

37. Core Team R. R: A language and environment for statistical computing. R foundation for statistical computing. 2013.

38. Leopoldo S, Lorena B, Cinzia A, et al. Two subtypes of mucinous adenocarcinoma of the colorectum: Clinicopathological and genetic features. *Ann Surg Oncol*. 2008;15(5):1429-1439.

39. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform*. 2005;38(5):404-415.

40. Zhou X, Obuchowski NA, McClish DK. Chapter 2. Measures of Diagnostic Accuracy. In: *Statistical Methods in Diagnostic Medicine.* 2nd ed. Hoboken: Wiley; 2011:13-57.

41. Søreide K. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *J Clin Pathol*. 2008;62(1):1.

42. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.

43. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatr*. 2011;48(4):277-287.

44. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39(4):561.

45. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77.

46. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762-775.

47. Kinsella RJ, Kähäri A, Haider S, et al. Ensembl BioMarts: A hub for data retrieval across taxonomic space. *Database*. 2011;2011:bar030.

48. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42(D1):D749-D755.

49. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797.

50. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: The NCBI database of genetic variation. *Nucleic Acids Res.* 2001;29(1):308-311.

51. Kent WJ, Sugnet C,W., Furey TS, et al. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996-1006.

52. Enomoto A, Wempe MF, Tsuchida H, et al. Molecular identification of a novel carnitine transporter specific to human testis: insights into the mechanism of carnitine recognition. *J Biol Chem*. 2002;277(39):36262-36271.

53. Aouida M, Poulin R, Ramotar D. The human carnitine transporter SLC22A16 mediates high affinity uptake of the anticancer polyamine analogue bleomycin-A5. *J Biol Chem*. 2009;285(9):6275-6284.

54. Fukuda T, Sugita S, Inatome R, Yanagi S. CAMDI, a novel disrupted in schizophrenia 1 (DISC1)-binding protein, is required for radial migration. *J Biol Chem*. 2010;285(52):40554-40561.

55. Kayser G, Gerlach U, Walch A, et al. Numerical and structural centrosome aberrations are an early and stable event in the adenoma-carcinoma sequence of colorectal carcinomas. *Virchows Archiv*. 2005;447(1):61-65.

56. Ishida N, Kawakita M. Molecular physiology and pathology of the nucleotide sugar transporter family (SLC35). *Pflügers Archiv*. 2004;447(5):768-775.

57. Xie Z, Zhang H, Tsai W, et al. Zinc finger protein ZBTB20 is a key repressor of alpha-fetoprotein gene transcription in liver. *Proc Natl Acad Sci U S A*. 2008;105(31):10859-10864.

58. Zhao J, Ren K, Tang J. Zinc finger protein ZBTB20 promotes cell proliferation in non-small cell lung cancer through repression of FoxO1. *FEBS Lett*. 2014;588(24):4536-4542.

59. Wang Q, Tan Y, Ren Y, et al. Zinc finger protein ZBTB20 expression is increased in hepatocellular carcinoma and associated with poor prognosis. *BMC Cancer*. 2011;11(1):271.

60. Fan W, Koch CA, de Hoog CL, Fam NP, Moran MF. The exchange factor ras-GRF2 activates ras-dependent and rac-dependent mitogen-activated protein kinase pathways. *Curr Biol*. 1998;8(16):935-939.

61. Crespo P, Calvo F, Sanz-Moreno V. Ras and rho GTPases on the move: The RasGRF connection. *Bioarchitecture*. 2011;1(4):200-204.

62. Wendeler MW, Paccaud J, Hauri H. Role of Sec24 isoforms in selective export of membrane proteins from the endoplasmic reticulum. *EMBO Rep*. 2006;8(3):258-264.

63. Goldenring JR. A central role for vesicle trafficking in epithelial neoplasia: Intracellular highways to carcinogenesis. *Nat Rev Cancer*. 2013;13(11):813-820.

64. Raffaello A, De Stefani D, Sabbadin D, et al. The mitochondrial calcium uniporter is a multimer that can include a dominant-negative pore-forming subunit. *EMBO J*. 2013;32(17):2362-2376.

65. Duchen MR. Mitochondria and calcium: From cell signalling to cell death. *J Physiol (Lond)*. 2000;529:57-68.

66. Brown GR, Hem V, Katz KS, et al. Gene: A gene-centered information resource at NCBI. *Nucleic Acids Res*. 2014;43:D36-D42.

67. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: The HGNC resources in 2015. *Nucleic Acids Res*. 2014;43:D1079-D1085.

68. Yamamoto S, Mochizuki H, Hase K, et al. Assessment of clinicopathologic features of colorectal mucinous adenocarcinoma. *Am J Surg*. 1993;166(3):257-261.

69. Janssens V, Goris J, Van Hoof C. PP2A: The expected tumor suppressor. *Curr Opin Genet Dev*. 2005;15(1):34-41.

70. Cristóbal I, Rincón R, Manso R, et al. Hyperphosphorylation of PP2A in colorectal cancer and the potential therapeutic value showed by its forskolin-induced dephosphorylation and activation. *Biochim Biophys Acta*. 2014;1842(9):1823-1829.

71. Cristóbal I, Manso R, Rincón R, et al. Phosphorylated protein phosphatase 2A determines poor outcome in patients with metastatic colorectal cancer. *Br J Cancer*. 2014;111(4):756-762.

72. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311-321.

73. Gabrielli F, Tofanelli S. Molecular and functional evolution of human DHRS2 and DHRS4 duplicated genes. *Gene*. 2012;511(2):461-469.

Chapter 3: Manuscript – "A Genome-wide Association Study Identifies Single Nucleotide Polymorphisms (SNPs) Associated with Time-to-Metastasis in Colorectal Cancer"

A version of this manuscript has been prepared for submission for publishing in a peer reviewed journal. Supplementary data is provided in Appendix C.

3.1 Authors and affiliations

Michelle E. Penney[1], Patrick S. Parfrey[2], Sevtap Savas[1,3], Yildiz E. Yilmaz[1,2,4]

1. Discipline of Genetics, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada

2. Discipline of Medicine, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada

3. Discipline of Oncology, Faculty of Medicine, Memorial University of Newfoundland, St. John's, Canada

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

3.2 Co-authorship statement

Michelle E. Penney: Aided in the development of the study design, conducted the statistical analyses, interpreted the results, and drafted the manuscript.

Dr. Patrick S. Parfrey: Provided the patient characteristics, genome-wide SNP genotype, and disease outcome data investigated in this study.

Dr. Sevtap Savas: Aided in the interpretation of the results and revised the manuscript.

Dr. Yildiz Yilmaz: Proposed the research question, developed the study design, reviewed all results and their interpretation, supervised the study, and revised the manuscript.

3.3 Abstract

Background: Differentiating between patients who will experience metastasis within a short time and who will be long-term survivors without metastasis is a critical aim in healthcare. The microsatellite instability (MSI)-high tumor phenotype is such a differentiator in colorectal cancer, as patients with these tumors are unlikely to experience metastasis. Our aim in this study was to determine if germline genetic variations could further differentiate colorectal cancer patients based on the long-term risk and timing of metastasis.

Methods: The patient cohort consisted of 379 stage I-III colorectal cancer patients with microsatellite stable or MSI-low tumors. We performed univariable analysis on 810,622 common single nucleotide polymorphisms (SNPs) under different genetic models. Depending on the long-term metastasis-free survival probability estimates, we applied a mixture cure model, Cox proportional hazards regression model, or log-rank test. For SNPs reaching Bonferroni-corrected significance ($p < 6.2 \times 10^{-8}$) having valid genetic models, multivariable analysis adjusting for significant baseline characteristics was conducted.

Results: After adjusting for significant baseline characteristics, specific genotypes of ten polymorphisms were significantly associated with early metastasis. These polymorphisms are three intergenic SNPs, rs5749032 (p=1.28e-10), rs2327990 (p=9.59e-10), rs1145724 (p=3e-8), and seven SNPs within the non-coding sequences of three genes: *FHIT* (p=2.59e-9), *EPHB1* (p=8.23e-9), and *MIR7515* (p=4.87e-8). These genes have been previously implicated in advanced tumor progression or metastasis in cancer.

Conclusions: Our results suggest novel associations of specific genotypes of SNPs with early metastasis in colorectal cancer patients. These associations, once replicated in other patient cohorts, could assist in the development of personalized treatment strategies for colorectal cancer patients.

3.4 Introduction

Cancer is an important and increasing problem worldwide. From 1990 to 2013, incidence for almost every cancer increased between 9% and 217% [1,2]. It is also an important cause of global mortality with over 8 million deaths caused by this disease in 2012 [1]. Metastasis, specifically, is responsible for approximately 90% of all cancer deaths [3,4]. Lately, there have been great advances in the development of therapeutics and increased survival of metastatic cancer patients [5-7]. However, despite serious efforts to better control this disease, the 5-year survival rate for several metastatic cancers is less than 20% [8].

A major contributor to the global cancer burden is colorectal cancer. In 2012, this disease was the second most common cancer in males and third in females. In addition, colorectal cancer caused almost 700,000 deaths worldwide in 2012 [1]. As with other cancer types, a main cause of death in colorectal cancer is metastasis. Several factors have been identified to have prognostic importance in colorectal cancer, including the tumor stage and tumor microsatellite instability (MSI) status [9]. A systematic review published in 2005 [10] concluded that colorectal cancer patients with microsatellite instability-high (MSI-H) tumors have a significantly improved prognosis than patients with microsatellite instability-low (MSI-L) and microsatellite stable (MSS) tumors. Furthermore, there is a significantly lower incidence of metastasis in MSI-H colorectal tumors [11-14]. However, despite the identification of such factors, there is still significant variability in patient outcomes that may be further explained by germline genetic variation. Following this hypothesis, genome-wide association studies (GWAS) can be conducted to identify such variations.

Routine research approaches in survival GWAS can be improved by considering a more focused study design and applying improved analysis methods to adequately and correctly address the research problem. By applying a focused study design, we could decrease the genetic and phenotypic heterogeneity, and hence, increase the statistical power of the association tests [15]. Among the components of a study design that can influence the study power are concentrating on a homogenous patient cohort and using a more explicit disease event phenotype. Homogeneous patient subgroups, by definition, consist of individuals that have a certain level of similarity. Such cohorts can be obtained, for example, by focusing on cancer subtypes that possess certain characteristics relevant to the research question [16]. This helps to reduce the variability in the disease trait and, thus, the association tests can be more powerful. In addition, when defining the time-to-event phenotype in survival analysis, it is important to use a specifically-defined endpoint [17]. Some endpoint definitions, such as death due to any cause, can include several different events. Although the number of events increases when such an endpoint definition is used, the ambiguity in the endpoint definition may lead to increased variability in the time-to-event phenotype and less informative results. It can be useful, instead, to consider a precisely defined endpoint that is an event specific to the disease being investigated, such as metastasis in cancer.

When analyzing metastasis as a clinical outcome, it is possible that not all patients will experience the outcome despite long follow-up times [18-21]. Consequently, the long-term metastasis-free survival probability estimates for these patient groups could plateau at a non-zero value. This indicates the patient cohort consists of a mixture of long-term metastasis-free survivors as well as patients who are susceptible to metastasis within the follow-up time [18,19,22-24]. Such a patient cohort can be properly investigated using the mixture cure model [17-19,21,24-26]. In fact, this model can determine two features for each potential predictive variable: 1) it can identify variables that are able to differentiate between patients who are susceptible to develop metastasis and who will potentially remain metastasis-free in the long-term, and 2) identify variables associated with time-to-metastasis in the susceptible patient sub-group. This model can make these determinations separately but simultaneously for each variable category. In addition, to make inferences on the long-term prognosis of a patient, long follow-up times are fundamentally important when modeling survival times with a mixture cure model [24]. However, when investigating high-dimensional data (such as genome-wide genotype data) using a mixture cure model, it is inevitable that the long-term metastasis-free survival probability estimates for a category of some variables will not plateau at a non-zero probability. For such variables, conventional survival models can be used to appropriately analyze the data.

Our specific objective in this study was to identify common single nucleotide polymorphisms (SNPs) that are associated with the long-term risk and timing of metastasis of colorectal cancer using a genome-wide genotype dataset. We focus our efforts on a subgroup of colorectal cancer patients with stage I-III MSI-L/MSS tumors. This study represents the first comprehensive study that aimed to identify the genetic markers that may be associated with the development of metastasis in colorectal cancer.

3.5 Methods

3.5.1 Ethics statement

Ethics approval for this study was obtained from the Health Research Ethics Board (HREB; #15.043). Since this study uses secondary data, patient consent specifically for this study was not required.

3.5.2 Patient cohort and genotype data

The patient cohort included in this study is a sub-cohort of the Newfoundland Colorectal Cancer Registry (NFCCR). The characteristics of the NFCCR cohort have been described previously [27,28]. NFCCR sought consent from participants; if the patient was deceased, consent was sought from a close relative [29]. These patients were followed until April 2010 [30].

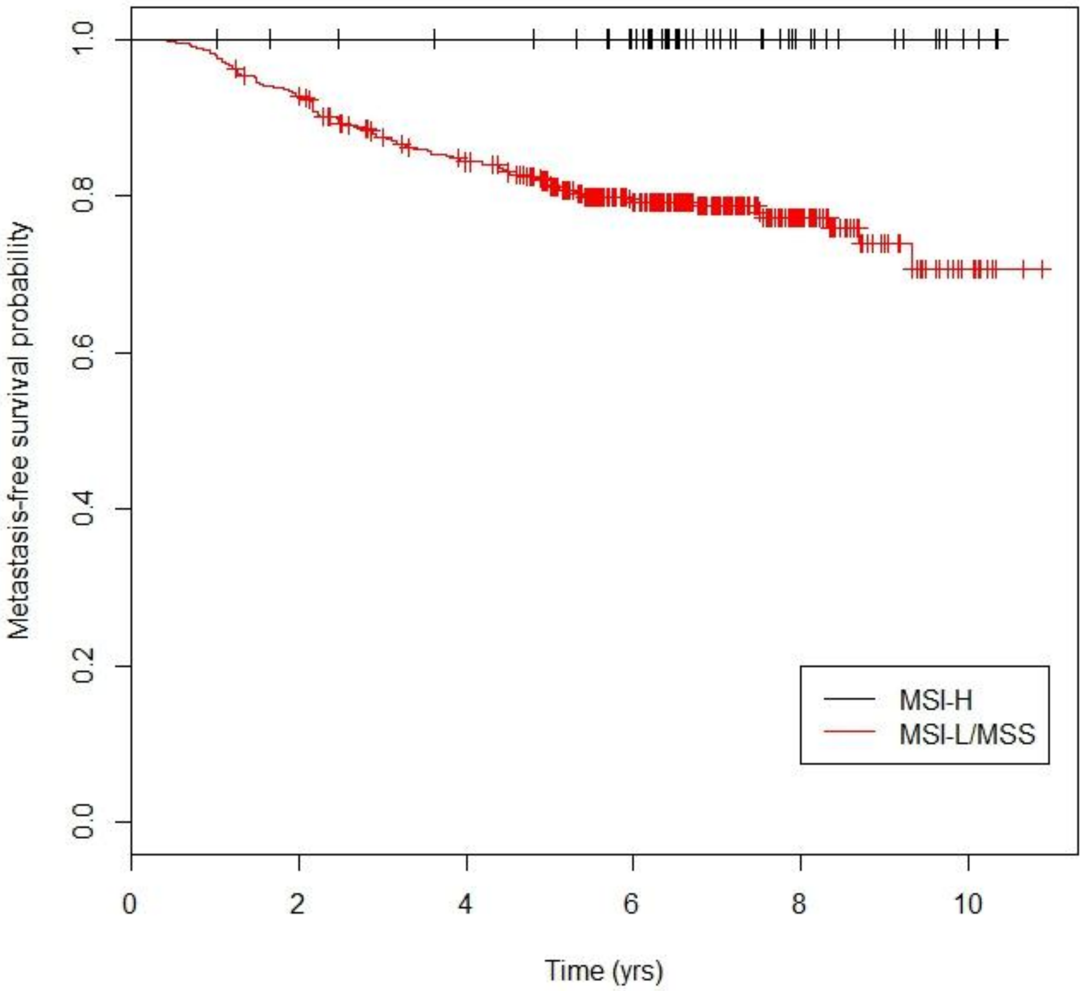
Genetic data was obtained from the NFCCR. Sample quality control steps on the genotype data were previously described by Xu et al. (2015) for another genome-wide survival study [31]. In short, germline DNA extracted from blood was available for 539 of the NFCCR patients. These DNA samples were subject to whole-genome SNP genotyping using the Illumina Omni-1 Quad human SNP genotyping platform at an outsourced company (Centrillion Bioscience, USA). Patients with discordant sex information, accidental duplicates, divergent or non-Caucasian ancestry, and first, second, or third-degree relatives were removed from the sample cohort [31]. There were 505 patients remaining in the quality-controlled data. Please note that no SNPs were removed due to high LD. In the previous genome-wide SNP-survival study [31], these 505 patients were examined in sub-groups (stage I-IV colorectal cancer patients with MSI-L/MSS tumors, and stage I-IV colon and rectal cancer patients regardless of their tumor MSI status) investigating associations between overall and disease-free survival times and genetic polymorphisms with a minor allele frequency (MAF) of at least 5%. The present study differs from this previous study in terms of the outcome of interest examined, minor allele frequencies of the genetic variants, and patients included, as explained below.

Further exclusion criteria/quality control measures were applied in order to address the objectives of this study. SNPs whose frequencies deviated from Hardy-Weinberg equilibrium, SNPs that had $>$ 5% missing values, and rare SNPs (minor allele frequency [MAF] $<$ 1%) were excluded, leaving 810,622 common SNPs. In contrast to the previous genome-wide survival study [31], the present study only considered stage I-III patients since patients with stage IV tumors (n=50) already have metastatic cancer. In addition, we focus our efforts on the MSI-L/MSS subgroup. This was motivated by the survival pattern observed when stratifying based on MSI status (excluding 20 patients with missing or unknown MSI tumor status and four patients due to lack of disease recurrence data). In the quality controlled patient data of stage I-III patients, there are no occurrences of metastasis in patients with MSI-H tumors (Figure 3.1). For this reason, 52 patients with MSI-H tumors have also been excluded from this study. The final study cohort consisted of 379 stage I-III patients with MSI-L/MSS tumors. Of these 379 patients, 21% experienced metastasis. The median follow-up time for metastasis was 6.3 years with the longest follow-up time being 10.9 years.

3.5.3 Data analysis

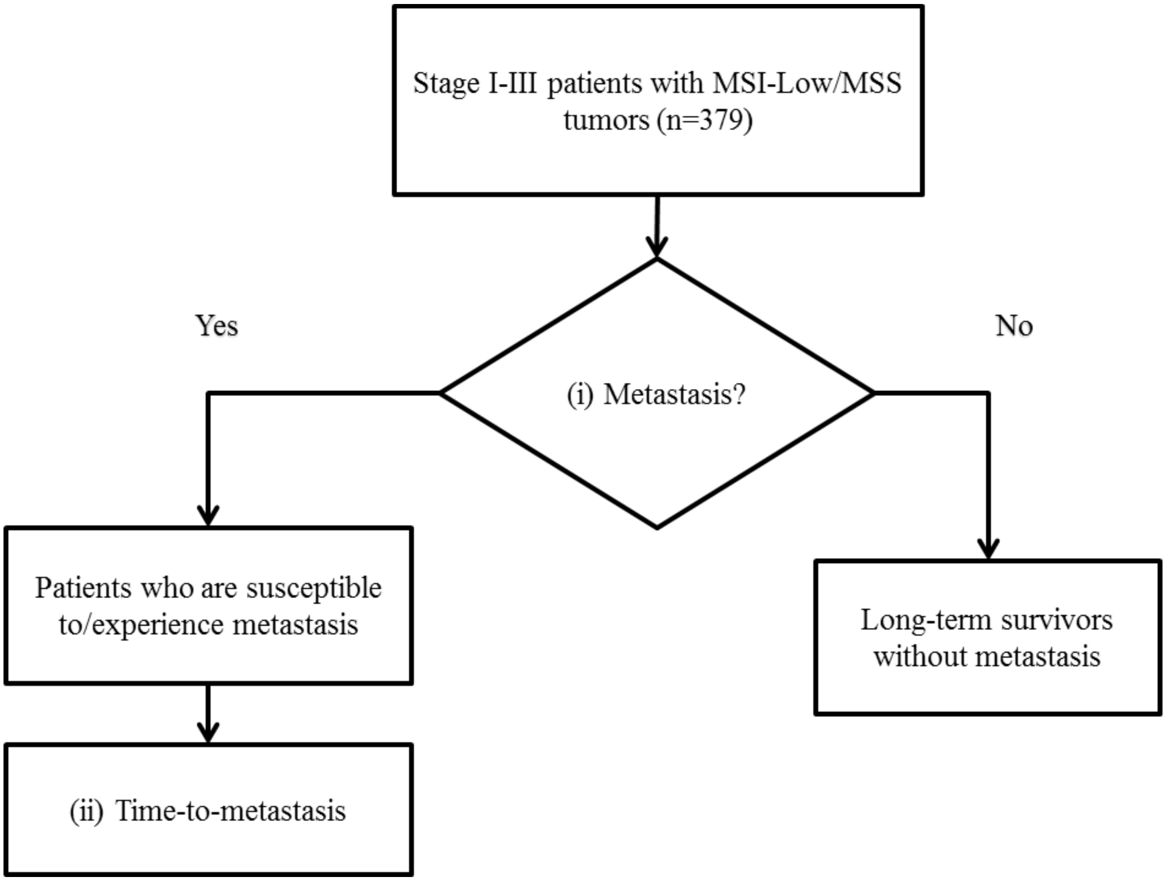
The survival outcome of interest throughout the analysis was time-to-metastasis. Patients who did not experience metastasis by the end of the follow-up time were censored at the time of the last follow-up. As seen in Figure 3.1, the long-term metastasis-free survival probability estimate for the patient subgroup with MSI-L/MSS tumors plateaus at 0.71 after being followed for just over 9 years. Since there is a plateau at a non-zero probability estimate, the mixture cure model is a potentially appropriate model to analyze this patient group [18,19,21,24-26]. In genetic association analyses such as the present one, this model can identify novel genes or regulatory regions, through the identification of SNPs, that are associated with (i) being a long-term survivor without metastasis and (ii) the time-to-metastasis in patients who are susceptible to/experience metastasis after diagnosis in the patient sub-cohort described above (Figure 3.2). As

Figure 3.1 Kaplan-Meier survival functions stratified by microsatellite instability (MSI) status



Kaplan-Meier survival functions stratified according to MSI status for the sub-cohort excluding stage IV patients and patients with unknown MSI tumor data (n=431). MSI-H: microsatellite instability high; MSI-L: microsatellite instability low; MSS: microsatellite stable.

Figure 3.2 Flow chart illustrating the aim of this study

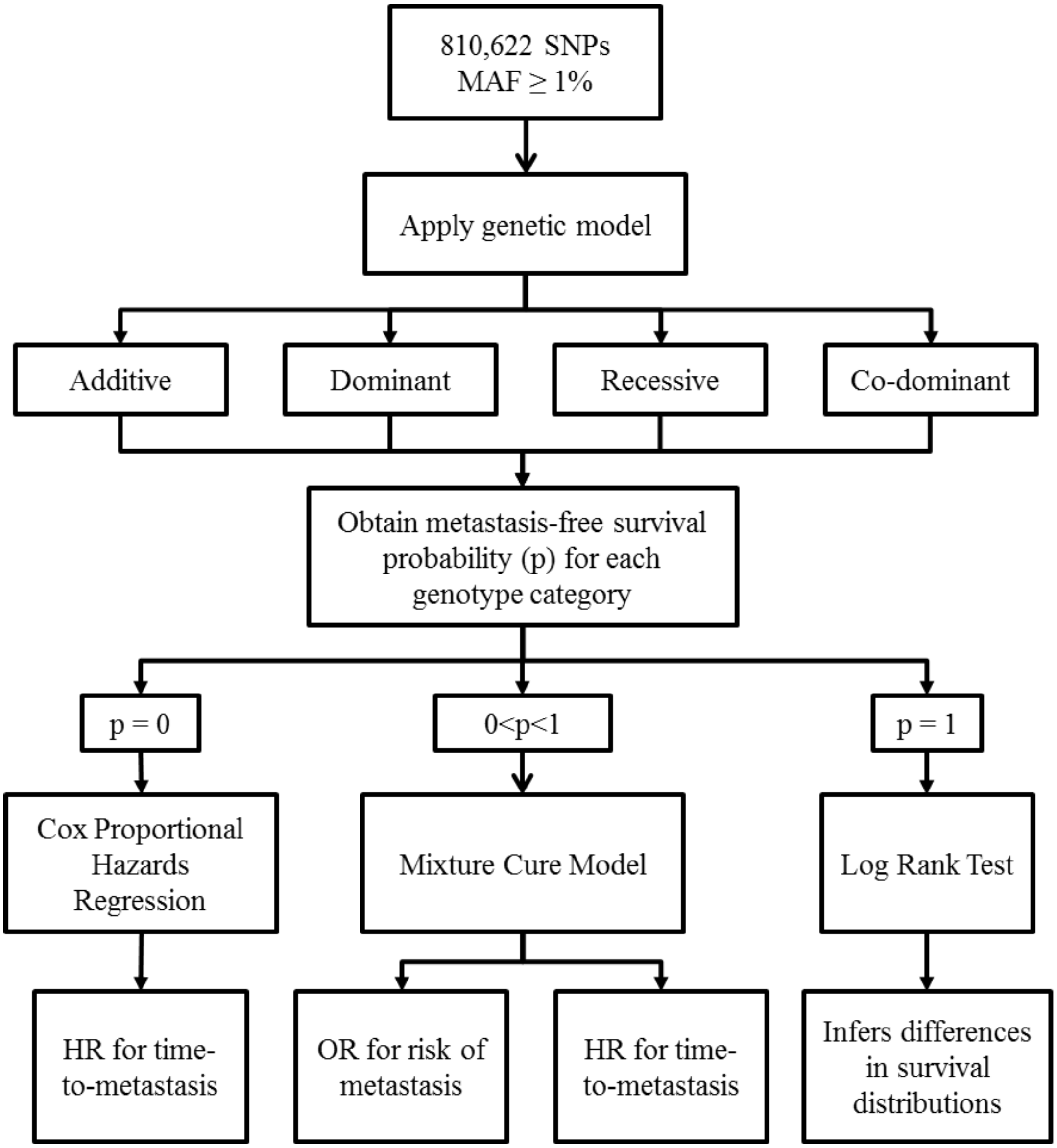

Stage I-III patients with MSI-Low/MSS tumors (n=379)
(i) Metastasis?
Yes
No
Patients who are susceptible to/experience metastasis
(ii) Time-to-metastasis
Long-term survivors without metastasis

mentioned earlier, both associations can be estimated using the mixture cure model.

3.5.3.1 Univariable analysis

Univariable analysis was performed on genome-wide SNP genotype data. This investigation required a detailed and comprehensive statistical analysis (Figure 3.3).

Figure 3.3 Methods of analysis used in this project


810,622 SNPs
MAF ≥ 1%
Apply genetic model
Additive
Dominant
Recessive
Co-dominant
Obtain metastasis-free survival probability (p) for each genotype category
p = 0
0<p<1
p = 1
Cox Proportional Hazards Regression
Mixture Cure Model
Log Rank Test
HR for time-to-metastasis
OR for risk of metastasis
HR for time-to-metastasis
Infers differences in survival distributions

Each SNP was analyzed under each genetic model using one of the three statistical methods listed.
HR: hazard ratio; MAF: minor allele frequency; OR: odds ratio; p: metastasis-free survival probability

First, for each SNP, all four genetic models were considered: additive, dominant, recessive, and co-dominant. However, for some SNPs, the number of patients with a genotype category was zero or very small (<2 patients) when the recessive (for 64,809 SNPs) and co-dominant (for 75,912 SNPs) genetic models were applied. As such, these SNPs were not analyzed under these specific genetic models.

For each SNP under a given genetic model, in order to determine if the mixture cure model was an appropriate model, we obtained the Kaplan Meier metastasis-free survival probability estimates at the end of the long-term follow-up time for each genotype category. If the long-term metastasis-free survival probability estimate was between zero and one, the mixture cure model was used. If the long-term metastasis-free survival probability estimate was zero for a genotype category, we applied the Cox proportional hazards regression model instead of the mixture cure model for the corresponding genetic model. For each significantly associated SNP identified under the Cox proportional hazards regression model, the proportionality assumption was assessed through a score test [32]. If the long-term metastasis-free survival probability estimate was one for a genotype category (i.e. if there is no metastasis within a given subgroup), we applied the log-rank test rather than fitting the mixture cure model or Cox proportional hazards regression model under the corresponding genetic model. In other words, SNPs that are associated with the probability of being a long-term metastasis-free survivor and the time-to-metastasis in patients who are susceptible to metastasis after diagnosis can be identified using the mixture cure model. For SNPs analyzed using the Cox proportional hazards regression model, we could test associations between specific genotype categories and time-to-metastasis. Finally, using the log-rank test, we could determine if there was a significant difference in the survival probability estimates between specified genotype categories.

All four genetic models were considered under the mixture cure model and for the log-rank test. However, only the recessive and co-dominant genetic models were used under the Cox proportional hazards regression model since there were no SNPs under the additive or dominant genetic models with corresponding genotypes yielding 0% metastasis-free survival estimate.

For this analysis, a Bonferroni-corrected p-value of 6.2×10^{-8} was required for significance.

3.5.3.2 Validity of the genetic model

For each significantly associated SNP, we assessed the fit of the corresponding genetic model under which it was identified. Since recessive, dominant, and additive models are nested models of the co-dominant model [33], we compared the results of the identified genetic model to the results of the co-dominant model using multiple approaches. We performed likelihood ratio tests to assess whether the identified genetic model was the plausible model. Additionally, for a sensitivity check, we constructed Kaplan-Meier curves under the co-dominant model and then checked whether or not the patterns of the curves for each genotype category were consistent with the estimated effects obtained under the identified genetic model. We also compared the coefficient estimates obtained in the identified additive, dominant, or recessive genetic model with the results obtained from the co-dominant model. This was done under the mixture cure model and the Cox proportional hazards regression model. Significant SNPs considered in the multivariable analysis are the ones identified in their most plausible genetic model. As given in the Results section, we did not identify any significantly associated SNP using the log-rank test, and hence validity of the genetic models was not assessed for this test.

3.5.3.3 Selection of significant baseline characteristics

Univariable analysis was also performed on the baseline characteristics to identify potential confounding factors to be adjusted for in the multivariable analysis. Patients with missing or unknown values for the baseline characteristics were excluded from this analysis. As such, we included only patients for which we had all data for the given baseline variable. This analysis was performed using the mixture cure model and the Cox proportional hazards regression model to select significant baseline characteristics for each model separately. Under both models, characteristics with a p-value of < 0.1 in the univariable analysis were prioritized and included in a backwards stepwise variable selection method to identify the final model with significant baseline characteristics. After this step, the significant baseline characteristics in the mixture cure model were tumor location, 5-fluorouracil (5-FU) treatment status, and stage (Supplementary Table S1). In the Cox proportional hazards regression model, the significant baseline characteristics were tumor location, stage, and *BRAF* V600E mutation status. In addition, although insignificant in the stepwise selection, 5-FU treatment status was forced into the model

(Supplementary Table S2). Of the significant baseline characteristics, only 5-FU treatment status and *BRAF* V600E mutation status had patients with missing or unknown values and were excluded from both models, resulting in 349 patients.

3.5.3.4 Multivariable analyses

Multivariable analysis was performed using the mixture cure model and the Cox proportional hazards regression model on significant SNPs identified in the respective univariable analyses adjusting for the selected baseline characteristics. These models were fitted using the genetic model in which the SNP was identified. As with the univariable analysis, a Bonferroni-corrected p-value of 6.2×10^{-8} was deemed significant.

All statistical analyses were conducted using R v 3.1.3 [34].

3.5.3.5 Bioinformatics analyses

The information related to functional effects of the significant SNPs was obtained using the Ensembl v. GRCh37.p13 Variant Effect Predictor (http://grch37.ensembl.org/Tools/VEP) [35]. Potential regulatory consequences of the significant SNPs were also explored through the RegulomeDB database (http://www.regulomedb.org/) [36]. Investigation of the genomic regions in which the SNPs were identified, including the identification of genes near intergenic SNPs, was performed using the UCSC Genome Browser (https://genome.ucsc.edu/) using the Human GRCh37/hg19 genome coordinates [37].

3.6 Results

The baseline characteristics of the patient cohort can be found in Table 3.1. The characteristics of the patient cohort considered in this study with genotype data (n=379) were comparable to the larger NFCCR cohort excluding stage IV and MSI-H tumors (n=517) (Supplementary Table S3). One-fifth (21%) of the patients in this cohort experienced metastasis within the follow-up time. There were noticeably more male (63%) than females (37%) in the patient cohort, but the proportions experiencing metastasis were similar (22% males; 21% females). Just over half the patients were treated with a 5-FU based treatment (57%), of which 28% experienced metastasis. More patients had tumors located in the colon (62%) than the rectum (38%), but a higher proportion of patients with rectal tumors experienced metastasis (27% vs. 18%). As expected, a higher proportion of stage III patients experienced metastasis (31%) within the follow-up time compared to stage I (10%) and II (19%). Most patients had non-mucinous tumors (91%) and one-fifth of these patients experienced metastasis.

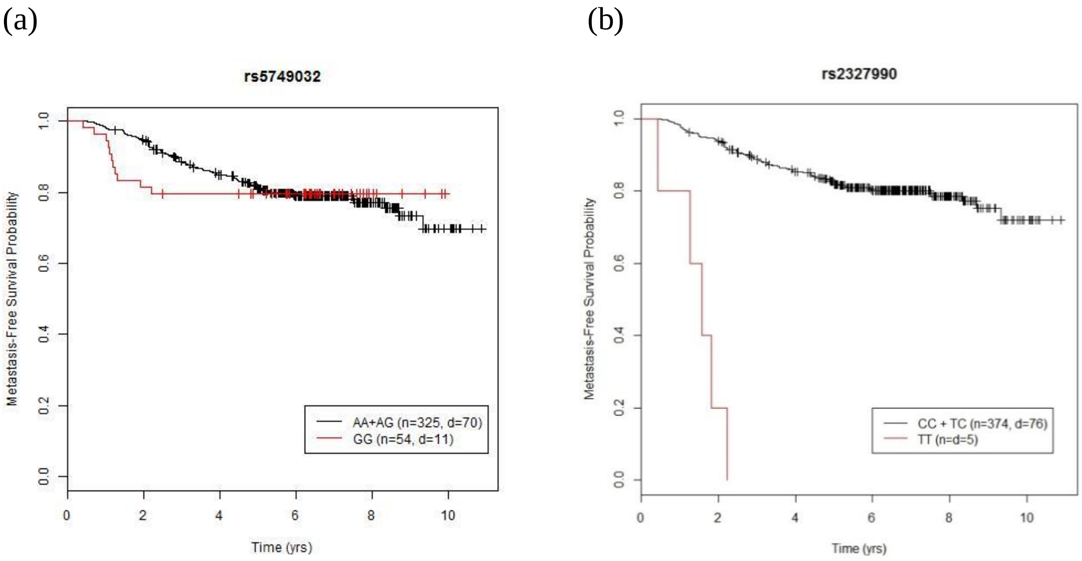
Using the univariable mixture cure model, we identified specific genotypes of nine SNPs that were significantly associated with time-to-metastasis (Figure 3.4, Supplementary Table S4, Supplementary Figure S1). These SNPs were identified under the dominant or recessive genetic model and did satisfy the test for genetic model validity. The nine significant SNPs were analyzed using a multivariable mixture cure model adjusting for significant baseline characteristics (Table 3.2, Supplementary Tables S5-S13). Of these, association of the minor allele homozygous genotype (genotype frequency=14%) in one SNP remained significant with time-to-metastasis in

Table 3.1 Baseline characteristics of the patient cohort (n=379) including metastasis proportions

Variable		**Number of patients**[a]	**% total**	**Number with metastasis**	**% metastasis**
Sex	Female	139	36.7%	29	20.9%
	Male	240	63.3%	52	21.7%
Age	≤60	157	41.4%	41	26.1%
	60-70	154	40.6%	29	18.8%
	>70	68	17.9%	11	16.2%
Familial risk	Low	196	51.7%	34	17.3%
	Intermediate/High	183	48.3%	47	25.7%
5-FU based treatment	5-FU treated	214	56.5%	59	27.6%
	Other/No chemo	159	42.0%	17	10.7%
	Unknown	6	1.6%	5	83.3%
Stage	I	81	21.4%	8	9.9%
	II	158	41.7%	30	19.0%
	III	140	36.9%	43	30.7%
Location	Colon	233	61.5%	41	17.6%
	Rectum	146	38.5%	40	27.4%
Histology	Non-mucinous	343	90.5%	75	21.9%
	Mucinous	36	9.5%	6	16.7%
Vascular invasion	Absence	242	63.9%	45	18.6%
	Presence	111	29.3%	30	27.0%
	Unknown	26	6.9%	6	23.1%
Lymphatic invasion	Absence	237	62.5%	44	18.6%
	Presence	116	30.6%	31	26.7%
	Unknown	26	6.9%	6	23.1%
BRAF V600E mutation	Absence	333	87.9%	72	21.6%
	Presence	19	5.0%	8	42.1%
	Unknown	27	7.1%	1	3.7%

a. Patients with MSI-H tumors and Stage IV patients were excluded. 5-FU: 5-fluorouracil

Figure 3.4 Kaplan-Meier survival function for the most significant SNPs in the multivariable analysis under the (a) mixture cure model and (b) Cox proportional hazards regression model



n: number of patients in that genotype category; d: number of metastasis in that genotype category.

(a) rs5749032 was the only SNP maintaining genome-wide significance after the multivariable analysis using the mixture cure model. In the rs5749032 GG genotype subgroup, the clear plateau at approximately 80% metastasis-free survival probability indicates the existence of a large proportion of long-term metastasis-free survivors.

(b) In the rs2327990 TT genotype subgroup, all the patients experienced metastasis within approximately the first two years. Therefore, a standard survival analysis method is appropriate.

the multivariable model (GG genotype of rs5749032; HR=15.86 [95% CI: 6.83-36.83], p=1.28E-10). We also obtained significant SNPs under the additive model. However, upon checking the validity of the genetic model, we found that the additive genetic model was not plausible for those SNPs. Thus, these results are not reported here.

Table 3.2 Results from the multivariable* analysis using the mixture cure model on the significant SNPs identified by the univariable mixture cure model

				Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Genomic location	**Genetic model**	**rs number (genotypes *a* vs. *b*)**	**Genotype freq.**	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
22:17793969	Recessive	rs5749032 (GG vs. AA + AG)	14%	0.38	0.14-1.07	0.066	15.86	6.83-36.83	1.28E-10
17:77361176	Co-Dominant	rs12949587 (CT vs. CC)	20%	0.66	0.32-1.37	0.261	7.56	3.44-16.61	4.63E-07
20:15111138	Co-Dominant	rs6110524 (AG vs. GG)	17%	0.95	0.44-2.04	0.887	4.80	2.00-11.53	4.52E-04
7:33913404	Recessive	rs3815652 (TT vs. CC + CT)	4%	0.59	0.13-2.65	0.488	12.97	3.26-51.66	2.78E-04
14:100691178	Recessive	rs756055 (CC vs. TT + TC)	13%	0.28	0.10-0.82	0.020	7.58	2.53-22.65	2.90E-04
14:100730920	Recessive	rs7153665 (AA vs. GG + AG)	13%	0.28	0.10-0.82	0.020	7.58	2.53-22.65	2.90E-04
11:100430053	Recessive	rs4754687 (AA vs. CC + CA)	11%	0.51	0.18-1.43	0.201	8.13	2.59-25.53	3.28E-04
5:155345221	Dominant	rs2163746 (CT + CC vs. TT)	24%	0.49	0.23-1.07	0.075	9.65	3.67-25.37	4.29E-06
5:155361116	Dominant	rs17053011 (TG + TT vs. GG)	24%	0.49	0.23-1.07	0.075	9.65	3.67-25.37	4.29E-06

*Adjusted for the significant baseline characteristics: tumor location, 5-fluorouracil treatment status, and tumor stage. Each SNP was analyzed separately adjusting for these factors. Linkage disequilibrium (LD) calculations indicated that rs756055 and rs7153665 as well as rs2163746 and rs17053011 are in complete pairwise LD ($r^2 = 1$). The SNPs listed yielded similar hazard ratio estimates under the univariable (Supplementary Table S4) and multivariable analyses. Consequently, all of the SNPs identified in this study could be considered independent prognostic factors for time-to-metastasis in colorectal cancer if the results are replicated using independent cohort data. Genotype freq.: frequency of genotype *a* calculated from the patient cohort; OR: odds ratio for metastasis comparing odds of metastasis in subgroup *a* with that in subgroup *b*; HR: hazard ratio comparing metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis; CI: confidence interval.

Univariable analysis under the Cox proportional hazards regression model identified 25 SNPs that were significantly associated with time-to-metastasis under the recessive and the co-dominant genetic models (Figure 3.4, Supplementary Table S14). The fitted genetic models were found to be the most plausible genetic model for each SNP. In addition, the proportionality assumption of the Cox proportional hazards regression model was not rejected for any of the significant SNPs. After adjusting for the significant baseline characteristics in the multivariable analysis, specific genotypes of nine SNPs remained significantly associated with time-to-metastasis (Table 3.3, Supplementary Tables S15-S23).

Last, of the SNPs analyzed in this study, there were no associations with the risk of metastasis reaching Bonferroni-corrected significance using either the mixture cure model or the log-rank test. However, promising associations were detected in the mixture cure model and are reported for interested readers (Table 3.4, Supplementary Figure S2).

3.7 Discussion

Distant metastasis is the most lethal event in colorectal cancer progression. Despite significant advances in treatment options, the 5-year survival rate for metastatic colorectal cancer patients is only 13.5% in the US [8]. Tumor MSI status is an important prognostic indicator in colorectal cancer, as patients with MSI-H tumors rarely experience metastasis [11-14]. Identifying additional biomarkers that can distinguish between patients

Table 3.3 Genotypes significantly associated with time-to-metastasis after adjusting for significant baseline characteristics identified in the Cox proportional hazards regression model

			Univariable			**Multivariable***		
Genomic location	**rs number (genotypes *a* vs. *b*)**	**Genotype freq.**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
20:16189263	rs2327990 (TT vs. CC + CT)	1.3%	21.97	8.42-57.33	2.74E-10	22.58	8.32-61.31	9.59E-10
3:134513356	rs11918092 (CC vs. AA + AC)	0.5%	216.98	35.64-1321.13	5.32E-09	535.33	63.20-4534.30	8.23E-09
3:134515336	rs3732568 (AA vs. CC + CA)	0.5%	216.98	35.64-1321.13	5.32E-09	535.33	63.20-4534.30	8.23E-09
3:59930672	rs2366964 (CC vs. TT + TC)	0.8%	41.19	11.81-143.66	5.40E-09	56.53	14.98-213.26	2.59E-09
2:6769988	rs1563948 (AA vs. GG + GA)	0.8%	34.43	10.35-114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
2:6773920	rs11694697 (TT vs. CC + CT)	0.8%	34.43	10.35-114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
2:6777992	rs11692570 (TT vs. CC + CT)	0.8%	34.43	10.35-114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
2:6779277	rs2219613 (TT vs. CC + CT)	0.8%	34.43	10.35-114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
6:91187510	rs1145724 (GG vs. AA + AG)	0.8%	30.76	9.27-102.03	2.14E-08	36.43	10.21-129.93	3.00E-08

*Adjusted for tumor location, 5-fluorouracil treatment status, *BRAF* V600E somatic mutation status, and tumor stage. Each SNP was analyzed separately adjusting for these factors. LD calculations indicated that rs11918092 and rs3732568 are in high pairwise LD ($r^2 = 0.96$). In addition, rs1563948, rs11694697, rs11692570, and rs2219613 are all highly linked to each other ($0.94 \leq r^2 \leq 1$). The SNPs listed yielded similar risk estimates under the univariable and multivariable analyses. Consequently, all of the SNPs identified in this study could be considered independent prognostic factors for time-to-metastasis in colorectal cancer if the results are replicated using independent cohort data. Genotype freq.: frequency of genotype *a* calculated from the patient cohort; HR: hazard ratio comparing metastasis rate in subgroup *a* with that in subgroup *b*; CI: confidence interval.

Table 3.4 Most significant associations with the risk of metastasis estimated in the univariable mixture cure model

Genomic location	**rs number (genotypes; *a* vs. *b*)**	**Genetic model**	**No.**	**Freq**	**MAF**	**Type of variant***	***Logistic regression model for metastasis probability***			***Proportional hazards model for time-to-metastasis***		
							OR	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
8:65783019	rs6985116 (CC vs. TT)	Co-Dominant	88	23%	48%	Intergenic	0.07	0.02-0.24	1.82E-05	2.61	0.68-10.07	0.1623
8:5438981	rs17354999 (AG & AA vs. GG)	Additive	252	66%	44%	Intergenic	2.93	1.79-4.79	1.98E-05	0.62	0.38-1.00	0.0524
5:105924416	rs10080115 (CT vs. TT)	Co-Dominant	149	39%	27%	Intergenic	0.24	0.13-0.47	2.03E-05	2.32	1.15-4.71	0.0195
22:47701711	rs4823630 (TC vs. CC)	Co-Dominant	173	46%	34%	Intergenic	0.24	0.12-0.47	2.58E-05	1.33	0.60-2.93	0.4844
8:5437805	rs1468386 (AA vs. GG)	Co-Dominant	94	25%	49%	Intergenic	7.41	2.87-19.17	3.61E-05	0.64	0.21-1.95	0.4298

*based on Ensembl database [38]

OR: odds ratio for metastasis comparing odds of metastasis in subgroup *a* with that in subgroup *b*.

HR: hazard ratio comparing metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis.

No.: number of patients with genotype *a*; MAF: minor allele frequency calculated from patient cohort analyzed; Freq: frequency of genotype *a* calculated from the patient cohort; CI: confidence interval.

who will experience metastasis in the short term and who will not experience metastasis in the long-term has clear clinical implications in the management and treatment of this disease. In this study, using a focused study design as well as applying appropriate and informative methods of analysis, we identified ten genetic polymorphisms significantly associated with time-to-metastasis in stage I-III colorectal cancer patients with MSI-L/MSS tumors after adjusting for significant baseline characteristics.

The mixture cure model identified a specific genotype (GG) of one SNP (rs5749032) that was significantly associated with early metastasis after adjusting for significant baseline characteristics (Table 3.2, HR=15.86, p=1.28e-10). This was a frequent genotype in the patient cohort (14%). Most patients with this genotype that experienced metastasis did so within the first 2 years post-diagnosis (Figure 3.4). After this time-point, patients with this genotype did not experience metastasis, despite the long-term follow-up for many patients. Essentially, this suggests that if metastasis occurs in patients with this genotype, it is likely to be in a relatively short time after diagnosis. A search in scientific literature and in the RegulomeDB database [36] did not return information about possible biological or regulatory functions of this polymorphism. However, according to the Haploreg database [39], there are no known SNPs in high linkage disequilibrium with this SNP. Thus, this polymorphism may have a direct biological effect on time-to-metastasis. The rs5749032 polymorphism is located in an intergenic sequence flanked by two genes: *CECR2* and *CECR3* (Supplementary Figure S3). *CECR2* is a protein-coding gene downstream of the SNP. The protein product is a transcription factor that is reported to be involved in chromatin remodeling [40] and may have an additional role in DNA damage response [41]. On the other hand, *CECR3* is a non-coding RNA, according to the Gene Entrez database [42]. At the present time, there are no reported relationships between these two genes and cancer. Finally, it is important to note that this association might not have been detected using the traditional survival analysis method of applying Cox proportional hazards regression model since the proportional hazards assumption was not satisfied (i.e. the survival curves cross; Figure 3.4) and there is a large proportion of long-term metastasis-free survivors (i.e. stable plateau at non-zero metastasis-free survival probability; Figure 3.4). We verified this by fitting a Cox proportional hazards regression model to this SNP. Under neither the univariable (HR=1.03 [95% CI: 0.54-1.94], p=0.93) nor the multivariable (HR=1.09 [95% CI: 0.57-2.10], p=0.80) Cox proportional hazards regression analysis was there a significant association. Overall, this SNP is a novel candidate biomarker deserving further investigations, particularly replicating its association and examining its potential biological link to metastasis.

For the SNPs with genotype categories showing 0% metastasis-free survival probability, the Cox proportional hazards regression model identified nine SNPs significantly associated with time-to-metastasis after adjusting for significant baseline characteristics (Table 3.3, Supplementary Figure S3). The most significant SNP, rs2327990, is an intergenic variant (Table 3.5). While there are no published reports about this SNP, according to the RegulomeDB database [36], there is some evidence that rs2327990 may affect the binding of transcription factors USF1 and USF2. The

Table 3.5 Variant information for the significant genotypes in the multivariable mixture cure and Cox proportional hazards regression models

Genomic location	**rs number (genotype[a])**	**MAF[b]**	**Type of variant (gene)[c]**	**DNA binding evidence[d]**
22:17793969	rs5749032 (GG)	40%	Intergenic	ND
20:16189263	rs2327990 (TT)	11%	Intergenic	Less likely to affect binding
3:134513356	rs11918092 (CC)	8%	Intronic (*EPHB1*)	Minimal binding evidence
3:134515336	rs3732568 (AA)	8%	Intronic (*EPHB1*)	Minimal binding evidence
3:59930672	rs2366964 (CC)	8%	Intronic (*FHIT*)	ND
2:6769988	rs1563948 (AA)	11%	Intronic (*MIR7515*)	Minimal binding evidence
2:6773920	rs11694697 (TT)	11%	Intronic (*MIR7515*)	ND
2:6777992	rs11692570 (TT)	11%	Intronic (*MIR7515*)	Minimal binding evidence
2:6779277	rs2219613 (TT)	11%	Intronic (*MIR7515*)	Minimal binding evidence
6:91187510	rs1145724 (GG)	9%	Intergenic	Minimal binding evidence

a. Risk increasing/decreasing genotype, b. Minor allele frequency (MAF) calculated from patient cohort analyzed. Values comparable to CEU population based on 1000 Genomes Project Phase 3 [43] data obtained through the Ensembl database (http://grch37.ensembl.org/), c. based on Ensembl database [38], d. based on RegulomeDB database [36].
ND: no data

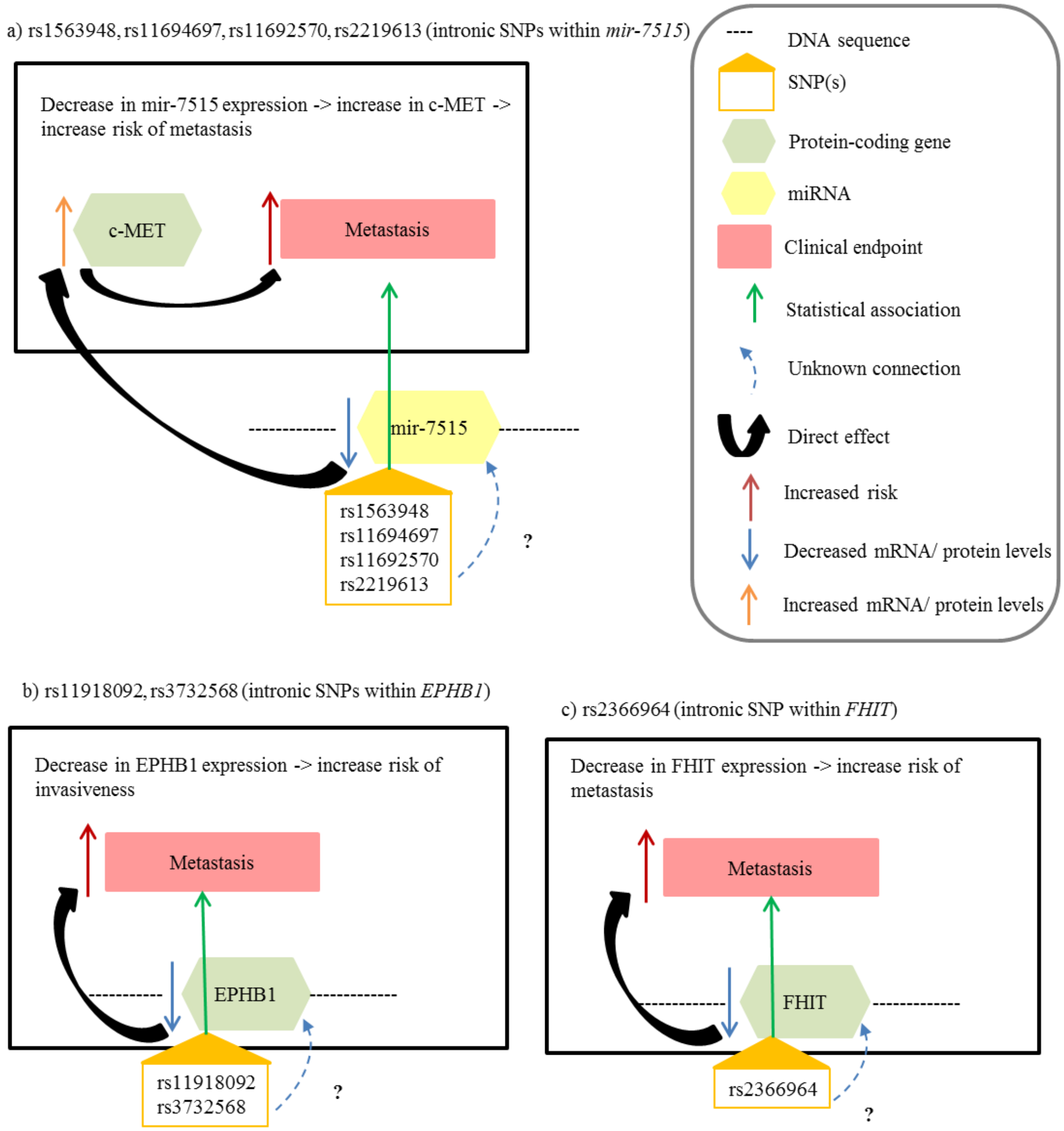
consequence of this potential regulatory function with regards to metastasis in colorectal cancer has yet to be investigated. This variant is located between a processed pseudogene, *PPIAP17*, and a protein coding gene, *KIF16B*. KIF16B is a kinesin-like protein that may be involved in intracellular trafficking [44]. While the function of *PPIAP17* is not known, there is a protein coding gene further upstream named *MACROD2*. This gene is quite interesting because one study examining 352 colorectal cancer patients identified *MACROD2* as the gene with the most prevalent and recurrent chromosomal breakpoints in colorectal tumors (41%) [45]. According to the Gene Entrez database [42], this gene encodes a deacetylase that removes ADP ribose from modified proteins. As also discussed by van den Broek et al. (2015), one of the target proteins of MACROD2 is GSK3β: active MACROD2 removes the mono-ADP-ribosyl units resulting in an increase in active GSK3β [46]. Interestingly, GSK3β is a regulator of the Wnt signaling pathway [47,48] and connections between upregulated Wnt signaling and distant metastasis in colorectal cancer have been identified [49,50]. Thus, when there is a reduction in active MACROD2 levels, this may lead to decreased GSK3β function, which in turn could lead to increased Wnt signaling and, accordingly, an increased risk of metastasis (Supplementary Figure S3*)*. Therefore, evaluating the presence of a link between the identified polymorphism, rs2327990, and MACROD2 expression levels and metastatic potential may prove to be valuable.

It is important to note that, although the quality control steps excluded rare SNPs (MAF $<$ 1%), when the recessive and co-dominant genetic models were applied to the raw genotype data, we obtained genotype frequencies that are rare in the patient cohort.

This is because these genetic models analyze the minor allele homozygous genotypes as one independent category. As a result, for the remaining eight significant SNPs reported from the Cox proportional hazards regression model, the genotype frequencies were less than 1% (Table 3.3). Consequently, although the associations were significant (possibly due to a high effect size [51]), the results may not be generalized to the population. Hence, we should interpret these results with caution. These SNPs were either intergenic (n=1; rs1145724; Supplementary Figure S3) or located within intronic sequences of three genes (n=7), including four linked SNPs in *MIR7515*, two linked SNPs in *EPHB1*, and one SNP in *FHIT* (Figure 3.5). There are no known functional consequences reported for these SNPs (Table 3.5) and the potential biological effects of these SNPs on these genes or metastasis in colorectal cancer are not presently known. However, the results of our study combined with previously published findings suggest that there may be potential relationships between these genes and metastasis in colorectal cancer. For example, low levels of FHIT [52,53] and increased levels of a target of *MIR7515*, c-MET [54], have been linked to increased risk of metastasis of colorectal tumors [55,56]. In addition, a reduced level of EPHB1 in colorectal cancer cells was associated with increased invasive potential in one study [57].

Additionally, although the univariable mixture cure model identified several SNPs significantly associated with time-to-metastasis in colorectal cancer, it did not detect any SNPs that were significantly associated with the long-term risk of metastasis. However, we report five of the most significantly associated SNPs for interested researchers (Table 3.5). These SNPs were all located in intergenic regions and have no

Figure 3.5 Known and hypothesized relationships between the identified SNPs, genes on either side of the SNPs, and the risk of metastasis



a) rs1563948, rs11694697, rs11692570, rs2219613 (intronic SNPs within *mir-7515*)

Decrease in mir-7515 expression -> increase in c-MET -> increase risk of metastasis

b) rs11918092, rs3732568 (intronic SNPs within *EPHB1*)

Decrease in EPHB1 expression -> increase risk of invasiveness

c) rs2366964 (intronic SNP within *FHIT*)

Decrease in FHIT expression -> increase risk of metastasis known regulatory consequences according to the RegulomeDB database [36]. The results contained odds ratio estimates different than 1 ($p < 3.7 \times 10^{-5}$), indicating that these SNPs could be differentiators for being long-term metastasis free survivors, but the associations did not reach the conservative Bonferroni-corrected significance level. This could be indicative of a lack of power due to the small number of patients who experienced metastasis. Consequently, these SNPs should be investigated in a larger cohort.

This is one of the first large-scale association studies that examined clinical outcomes in colorectal cancer. Two other studies published previously investigated the prognostic value of genome-wide genetic polymorphisms on colorectal cancer patient outcomes. As explained in the Methods, Xu et al. (2015) performed a genome-wide association study with the aim of identifying common genetic polymorphisms associated with overall and disease-free survival times in stage I-IV colorectal cancer patient cohorts [31]. This study did not identify associations reaching genome-wide significance levels. In addition, Phipps et al. (2016) investigated associations between genome-wide common genetic variants and survival outcomes in patients enrolled in six prospective cohort studies [58]. These authors also performed an analysis on a sub-group of their study cohort by focusing only on those patients who had already experienced metastasis at diagnosis (i.e. stage IV patients) and identified a set of SNPs in their pooled analysis that were significantly associated with overall survival times. In contrast to both of these studies, our study considered time-to-metastasis as the survival outcome, applied appropriate statistical methods due to the investigation of metastasis, and focused on patients with stage I-III and MSI-L/MSS tumors only. Thus, this study is different from both of these previous studies and brings a new depth into colorectal cancer research in terms of both its design and significant findings.

A large strength of this study is the comprehensive study design. We applied appropriate methods of analysis based on the endpoint of choice and the characteristics of the patient cohort subgroups we considered rather than applying the widely used Cox proportional hazards regression model only. In addition, by concentrating our efforts on a sub-cohort determined by the MSI tumor status and the tumor stage, we created a more homogeneous study cohort with an undifferentiated survival pattern (Figure 3.1). This enabled us to reduce the genetic and phenotypic variability in the cohort to identify potential prognostic biomarkers. This intricate study design allowed for a more powerful analysis although we had a moderate number of patients. We also applied four genetic models to ensure a complete and informative investigation. Finally, it is important to note that, in this study, we proposed and applied a framework for conducting a genome-wide association study of time-to-metastasis in curable cancer types. The study design and statistical methods utilized in this study are pertinent to any cancer type that has a large proportion of long-term metastasis-free survivors. This is significant, since advances in medical research are creating more patient cohorts with such a characteristic. Consequently, this study not only identified potential biomarkers for early metastasis in colorectal cancer patients, but also demonstrated an advanced and informative analysis approach to potentially enrich prognostic research in other cancer types.

In conclusion, this is the first study to investigate genetic associations with time-to-metastasis in colorectal cancer patients using such a large genetic data set and the first study where a mixture cure model was used for a high dimensional genetic data analysis. More importantly, for the first time, significant associations between genome-wide SNP genotype data and time-to-metastasis were detected in colorectal cancer patients. The identified genetic variations represent a novel set of SNPs and genes that may have biological roles in colorectal cancer progression and metastasis in these patients. Once replicated, these results could aid in providing a means to distinguish colorectal cancer patients who are at an increased risk of early metastasis, which could be valuable in the clinical care of these patients as well as contribute to individualized therapies.

3.8 Acknowledgements

We thank the patients and families that participated in NFCCR and all the NFCCR personnel/investigators who contributed to NFCCR. This work was primarily funded by the Ignite fund - Research and Development Corporation (RDC) Newfoundland and Labrador (NL) awarded to Y.E. Yilmaz. S. Savas is a Beatrice Hunter Cancer Research Institute (BHCRI) senior scientist. M. Penney is supported by a Translational and Personalized Medicine Initiative (TPMI) fellowship. The funding sources were not involved in the study design; in the collection, analysis or interpretation of data; in the writing of the report; or in the decision to submit the paper for publication.

3.9 References

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359-E386.

2. Global Burden of Disease Cancer Collaboration. The Global Burden of Cancer 2013. *JAMA Oncol*. 2015;1(4):505-527.

3. Monteiro J, Fodde R. Cancer stemness and metastasis: Therapeutic consequences and perspectives. *Eur J Cancer*. 2010;46(7):1198-1203.

4. Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. *Science*. 2011;331(6024):1559.

5. Hersh EM, Del Vecchio M, Brown MP, et al. A randomized, controlled phase III trial of nab-paclitaxel versus dacarbazine in chemotherapy-naïve patients with metastatic melanoma. *Ann Oncol*. 2015;26(11):2267-2274.

6. Hurwitz H, Fehrenbacher L, Novotny W, et al. Bevacizumab plus irinotecan, fluorouracil, and leucovorin for metastatic colorectal cancer. *N Engl J Med*. 2004;350(23):2335-2342.

7. Aitelhaj M, Lkhoyaali S, Rais G, Boutayeb S, Errihani H. First line chemotherapy plus trastuzumab in metastatic breast cancer HER2 positive - observational institutional study. *Pan Afr Med J*. 2014;24:324.

8. Howlader N, Noone A, Krapcho M, et al. SEER Cancer Statistics Review, 1975-2013. *National Cancer Institute, Bethesda, MD*. 2016; based on November 2015 SEER data submission (posted to the SEER web site, April 2016). https://seer.cancer.gov/csr/1975_2014/

9. Compton CC, Fielding LP, Burgart LJ, et al. Prognostic factors in colorectal cancer. *Arch Pathol Lab Med*. 2000;124(7):979-994.

10. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol*. 2005;23(3):609-618.

11. Buckowitz A, Knaebel H, Benner A, et al. Microsatellite instability in colorectal cancer is associated with local lymphocyte infiltration and low frequency of distant metastases. *Br J Cancer*. 2005;92(9):1746-1753.

12. Kloor M, Staffa L, Ahadova A, von KD. Clinical significance of microsatellite instability in colorectal cancer. *Langenbeck's Arch Surg*. 2014;399(1):23-31.

13. Malesci A, Laghi L, Bianchi P, et al. Reduced likelihood of metastases in patients with microsatellite-unstable colorectal cancer. *Clin Cancer Res*. 2007;13(13):3831.

14. Lim S, Jeong S, Lee MR, et al. Prognostic significance of microsatellite instability in sporadic colorectal cancer. *Int J Colorectal Dis*. 2004;19(6):533-537.

15. MacRae CA, Vasan RS. Next-generation genome-wide association studies. *Circ Cardiovasc Genet*. 2011;4(4):334.

16. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer*. 2005;5(2):142-149.

17. Lawless J. *Statistical models and methods for lifetime data* . 2nd Ed, Wiley Series in Probability and Statistics. Hoboken, N.J.: Wiley-Interscience; 2003.

18. Paoletti X, Asselain B. Survival analysis in clinical trials: Old tools or new techniques. *Surg Oncol*. 2010;19(2):55-58.

19. Yilmaz YE, Lawless JF, Andrulis IL, Bull SB. Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. *J Clin Oncol*. 2013;31(16):2047-2054.

20. Lawless JF, Yilmaz YE. Semiparametric estimation in copula models for bivariate sequential survival times. *Biom J*. 2011;53(5):779-796.

21. Forse C, Yilmaz Y, Pinnaduwage D, et al. Elevated expression of podocalyxin is associated with lymphatic invasion, basal-like phenotype, and clinical outcome in axillary lymph node-negative breast cancer. *Breast Cancer Res Treat*. 2013;137(3):709-719.

22. Lambert PC, Dickman PW, Weston CL, Thompson JR. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *J R Stat Soc Ser C*. 2010;59(1):35-55.

23. Tsodikov AD, Ibrahim JG, Jakovlev AY. Estimating cure rates from survival data. *J Am Stat Assoc*. 2003;98(464):1063-1078.

24. Sy JP, Taylor JMG. Estimation in a cox proportional hazards cure model. *Biometrics*. 2000;56(1):227-236.

25. Bejan-Angoulvant T, Bouvier A, Bossard N, et al. Hazard regression model and cure rate model in colon cancer relative survival trends: Are they telling the same story? *Eur J Epidemiol*. 2008;23(4):251-259.

26. Cox DR. Summary comments. *Surg Oncol*. 2010;19(2):61.

27. Woods MO, Younghusband HB, Parfrey PS, et al. The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut*. 2010;59(10):1369-1377.

28. Wish TA, Hyde AJ, Parfrey PS, et al. Increased cancer predisposition in family members of colorectal cancer patients harboring the p.V600E BRAF mutation: A population-based study. *Cancer Epidemiol Biomarkers Prevent*. 2010;19(7):1831-1839.

29. Green RC, Green JS, Buehler SK, et al. Very high incidence of familial colorectal cancer in newfoundland: A comparison with ontario and 13 other population-based studies. *Fam Cancer*. 2007;6(1):53-62.

30. Negandhi AA, Hyde A, Dicks E, 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*. 2013;8(4):e61469.

31. Xu W, Xu J, Shestopaloff K, et al. A genome wide association study on newfoundland colorectal cancer patients' survival outcomes. *Biomarker Res*. 2015;3(1):6.

32. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515-526.

33. Thomas DC. Chapter 4: Basic Epidemiologic and Statistic Principles. *Statistical methods in genetic epidemiology*. Cary: Oxford University Press; 2004.

34. Core Team R. R: A language and environment for statistical computing. R foundation for statistical computing. 2013.

35. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. *Genome Biol*. 2016;17:122.

36. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797.

37. Kent WJ, Sugnet C,W., Furey TS, et al. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996-1006.

38. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42(D1):D749-D755.

39. Ward LD, Kellis M. HaploReg: A resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Research*. 2012; 40(Database issue):D930-D934

40. Banting GS, Barak O, Ames TM, et al. CECR2, a protein involved in neurulation, forms a novel chromatin remodeling complex with SNF2L. *Hum Mol Genet*. 2005;14(4):513-524.

41. Lee S, Park E, Lee H, Lee YS, Kwon J. Genome-wide screen of human bromodomain-containing proteins identifies Cecr2 as a novel DNA damage response protein. *Mol Cells*. 2012;34(1):85-91.

42. Brown GR, Hem V, Katz KS, et al. Gene: A gene-centered information resource at NCBI. *Nucleic Acids Res*. 2014;43:D36-D42.

43. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

44. Hoepfner S, Severin F, Cabezas A, et al. Modulation of receptor recycling and degradation by the endosomal kinesin KIF16B. *Cell*. 2005;121(3):437-450.

45. van den Broek E, Dijkstra MJJ, Krijgsman O, et al. High prevalence and clinical relevance of genes affected by chromosomal breaks in colorectal cancer. *PLoS ONE*. 2015;10(9):e0138141.

46. Rosenthal F, Feijs KLH, Frugier E, et al. Macrodomain-containing proteins are new mono-ADP-ribosylhydrolases. *Nat Struct Mol Biol*. 2013;20(4):502-507.

47. Caspi M, Zilberberg A, Eldar-Finkelman H, Rosin-Arbesfeld R. Nuclear GSK-3β inhibits the canonical wnt signalling pathway in a β-catenin phosphorylation-independent manner. *Oncogene*. 2008;27(25):3546-3555.

48. Wu D, Pan W. GSK3: A multifaceted kinase in wnt signaling. *Trends Biochem Sci*. 2009;35(3):161-168.

49. Ormanns S, Neumann J, Horst D, Kirchner T, Jung A. WNT signaling and distant metastasis in colon cancer through transcriptional activity of nuclear Î²-catenin depend on active PI3K signaling. *Oncotarget*. 2014;5(10):2999-3011.

50. Sack U, Stein U. Wnt up your mind - intervention strategies for S100A4-induced metastasis in colon cancer. *Gen Physiol Biophys*. 2009;28:F55-64.

51. Konigorski S, Yilmaz YE, Pischon T. Comparison of single-marker and multi-marker tests in rare variant association studies of quantitative traits. *PLoS ONE*. 2017;12(5):e0178504.

52. Mimori K, Ishii H, Nagahara H, et al. FHIT is up-regulated by inflammatory stimuli and inhibits prostaglandin E_2-mediated cancer progression. *Cancer Res*. 2006;66(5):2683.

53. Mady HH, Melhem MF. FHIT protein expression and its relation to apoptosis, tumor histologic grade and prognosis in colorectal adenocarcinoma: An immunohistochemical and image analysis study. *Clin Exp Metastasis*. 2002;19(4):351-358.

54. Lee JM, Yoo JK, Yoo H, et al. The novel miR-7515 decreases the proliferation and migration of human lung cancer cells by targeting c-met. *Mol Cancer Res*. 2013;11(1):43.

55. Cui Y, Jiao H, Ye Y, et al. FOXC2 promotes colorectal cancer metastasis by directly targeting MET. *Oncogene*. 2015;34(33):4379-4390.

56. Elliott VA, Rychahou P, Zaytseva YY, Evers BM. Activation of c-met and upregulation of CD44 expression are associated with the metastatic phenotype in the colorectal cancer liver metastasis model. *PLoS ONE*. 2014;9(5):e97432.

57. Sheng Z, Wang J, Dong Y, et al. EphB1 is underexpressed in poorly differentiated colorectal cancers. *Pathobiology*. 2008;75(5):274-280.

58. Phipps AI, Passarelli MN, Chan AT, et al. Common genetic variation and survival after colorectal cancer diagnosis: A genome-wide analysis. *Carcinogenesis*. 2016;37(1):87-95.

Chapter 4: General Conclusions

Colorectal cancer is a serious health concern worldwide with significant impact on global cancer mortality [1]. Obtaining a better understanding of the potential genetic foundation of colorectal cancer subtype development and clinical characteristics is essential in improving patient care and outcomes. With this in mind, the projects in this dissertation aimed to identify genetic associations with a histological variant of colorectal cancer as well as with the risk and timing of metastasis in a Newfoundland colorectal cancer patient cohort.

The mucinous histological variant of colorectal cancer is a distinct tumor subtype that appears to exhibit aggressive tendencies. However, the genetic markers of this histological subtype are largely unknown. Previous studies investigating the genetic basis of mucinous tumor histology in colorectal cancer have been focused on candidate genes [241-243]. Consequently, in the manuscript presented in Chapter 2, we aimed to detect common and rare germline genetic variants that are associated with mucinous tumor histology in colorectal cancer patients using a genome-wide approach. This is the first time such a comprehensive investigation that included the analysis of a genome-wide set of both common and rare genetic variants has been performed related to colorectal cancer tumor histology. We performed single-SNP analysis to detect associations between common polymorphisms and mucinous tumor histology. In addition, we analyzed the rare variants using a recently developed multi-marker test procedure. As a result of these analyses, we identified novel polymorphisms (Table 2.2) and genes (Table 2.3) that may explain a portion of the diversity in colorectal tumor histology. In fact, there was a significant increase in the discriminatory accuracy of the model to differentiate between mucinous and non-mucinous tumors when the identified common SNPs were added to the model containing the significant baseline characteristics. Overall, the results from this study provide novel candidate biomarkers that may help ascertain the genetic basis of tumor histology in colorectal cancer patients.

Metastasis is the main cause of death in colorectal cancer [244,245]. However, it is possible that not all patients will experience metastasis, resulting in a patient population that is a mixture of individuals who are susceptible and non-susceptible to metastasis. Regrettably, predicting the risk of metastasis in stage I-III colorectal cancer patients remains difficult as there is still much unexplained variability in the long-term risk and timing of metastasis. There is one factor, the MSI-H tumor status, which is associated with extremely low incidence of metastasis (Figure 3.1). However, the remaining patients, those with MSI-L/MSS tumors, display an undifferentiated survival pattern. In light of this, the second manuscript of this thesis (Chapter 3) intended to illuminate some of the genetic basis of distant metastasis in colorectal cancer patients. To that end, we performed the first genome-wide association study which aimed to identify common germline genetic variations associated with the long-term risk and timing of metastasis in colorectal cancer patients with MSI-L/MSS tumors. After extensive analyses, we identified specific genotypes of ten polymorphisms that were significantly and independently associated with early metastasis. Of particular interest was a specific genotype of rs5749032, which was frequent in the patient cohort (14%) and provided the most significant association. Patients with this genotype that experienced metastasis did so within a short time after diagnosis, after which there were no instances of metastasis during the follow up and these patients could be considered statistically cured. The clinical implications of this result could be significant: for example, patients could be screened for this genotype and, if they have this genotype, could receive aggressive treatment and close monitoring for the first two years post-diagnosis. If patients with this genotype do not experience metastasis in the first two years, they are likely to be statistically cured and treatment/monitoring can be reduced or stopped. This may minimize the likelihood of over-treating the patient, as well as decrease the burden on the healthcare system.

Interestingly, the study described in Chapter 3 was the first to apply a mixture cure model in a genome-wide association study. When there is empirical evidence of a large proportion of long-term metastasis-free survivors in a patient cohort, the mixture cure model is appropriate to model time-to-metastasis. While such models are relatively well-known in statistical science, they have yet to be widely applied in medical research. With continuing advancements in treatment and prognostic research, it is expected that such mixed-survival cohorts will become more common and, accordingly, mixture cure models may be an important tool to properly model the time-to-event. Furthermore, by considering colorectal cancer patients with stage I-III MSI-L/MSS tumors, we decreased the phenotypic and genetic variability which helped isolate the potential effects of the polymorphisms. Furthermore, based on the long-term metastasis-free survival estimates, we applied appropriate statistical models and methods for each polymorphism. As a result of such methodological considerations, we were able to identify significant genetic associations. Therefore, I hope that the approach used in this project can act as a resource for other researchers and inspire them to investigate other diseases with a significant cure fraction to facilitate the continuation of novel discoveries in prognostic research.

In conclusion, this Master’s thesis presents two projects that were quite interdisciplinary in nature, including aspects of molecular genetics, epidemiology, and applied statistics. Once replicated, these results could strengthen our overall understanding of colorectal cancer biology as well as possibly assist in the development of personalized treatment strategies for colorectal cancer patients. While much work remains to be done in these areas of colorectal cancer genetics research, this thesis provides a set of new candidate polymorphisms and genes that may enhance our comprehension of the etiology of colorectal cancer tumor histology and distant metastasis.

REFERENCES

Note: These are references for Chapter 1 and Chapter 4 only. The references for Chapter 2 and Chapter 3 are given at the end of the respective chapters.

1. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359-E386.

2. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian Cancer Statistics 2017. *Toronto, ON: Canadian Cancer Society*. 2017. Available at: cancer.ca/Canadian-Cancer-Statistics-2017-EN.pdf

3. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet*. 2014;383(9927):1490-1502.

4. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. *N Engl J Med*. 2003;348(10):919-932.

5. Canadian Cancer Society. Risk factors for colorectal cancer. http://www.cancer.ca/en/cancer-information/cancer-type/colorectal/risks/?region=nl. Updated 2017. Accessed 1/9, 2017

6. Zhang K, Civan J, Mukherjee S, Patel F, Yang H. Genetic variations in colorectal cancer risk and clinical outcome. *World J Gastroenterol*. 2014;20(15):4167-4177.

7. American Society of Clinical Oncology Cancer.Net Editorial Board. Colorectal cancer - symptoms and signs. http://www.cancer.net/cancer-types/colorectal-cancer/symptoms-and-signs. Updated 2016. Accessed 1/9, 2017.

8. Haggar FA, Boushey RP. Colorectal cancer epidemiology: Incidence, mortality, survival, and risk factors. *Clin Colon Rectal Surg*. 2009;22(4):191-197.

9. Riihimäki M, Hemminki A, Sundquist J, Hemminki K. Patterns of metastasis in colon and rectal cancer. *Sci Rep*. 2016;6:29765.

10. Burt RW, Cannon JA, David DS, et al. Colorectal cancer screening. *J Natl Compr Cancer Netw*. 2013;11(12):1538-1575.

11. Burch JA, Soares-Weiser K, St John,D.J.B., et al. Diagnostic accuracy of faecal occult blood tests used in screening for colorectal cancer: A systematic review. *J Med Screen*. 2007;14(3):132-137.

12. Elmunzer BJ, Hayward RA, Schoenfeld PS, Saini SD, Deshpande A, Waljee AK. Effect of flexible sigmoidoscopy-based screening on incidence and mortality of colorectal cancer: A systematic review and meta-analysis of randomized controlled trials. *PLoS Medicine*. 2012;9(12):e1001352.

13. Screen Colons Canada. Screen colons canada. http://www.screencolons.ca/. Updated 2017. Accessed 2017.

14. Canadian Partnership Against Cancer. Cancer screening in Canada: An overview of screening participation for breast, cervical and colorectal cancer. Toronto: Canadian Partnership Against Cancer; January 2015.

15. Canadian Cancer Society. Screening for colorectal cancer. http://www.cancer.ca/en/prevention-and-screening/early-detection-and-screening/screening/screening-for-colorectal-cancer/?region=bc. Updated 2017. Accessed 2017.

16. Canadian Task Force on Preventive Health Care. Recommendations on screening for colorectal cancer in primary care. *CMAJ*. 2016;188(5):340-348.

17. Woods MO, Hyde AJ, Curtis FK, et al. High frequency of hereditary colorectal cancer in newfoundland likely involves novel susceptibility genes. *Clin Cancer Res*. 2005;11(19):6853.

18. Green RC, Green JS, Buehler SK, et al. Very high incidence of familial colorectal cancer in newfoundland: A comparison with ontario and 13 other population-based studies. *Fam Cancer*. 2007;6(1):53-62.

19. Eastern Health. Newfoundland and Labrador colon cancer screening program. St. John's: Eastern Health. 2015. http://www.easternhealth.ca/WebInWeb.aspx?d=3&id=1242&p=1078

20. Lynch HT, Smyrk TC, Watson P, et al. Genetics, natural history, tumor spectrum, and pathology of hereditary nonpolyposis colorectal cancer: An updated review. *Gastroenterology*. 1993;104(5):1535-1549.

21. Peltomäki P, Vasen H. Mutations associated with HNPCC predisposition - update of ICG-HNPCC/INSiGHT mutation database. *Dis Markers*. 2004;20(4-5):269-276.

22. Stoffel E, Mukherjee B, Raymond VM, et al. Calculation of risk of colorectal and endometrial cancer among patients with lynch syndrome. *Gastroenterology*. 2009;137(5):1621-1627.

23. Vasen HFA, Watson P, Mecklin J, Lynch HT. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, lynch syndrome) proposed by the international collaborative group on HNPCC. *Gastroenterology*. 1999;116(6):1453-1456.

24. Umar A, Boland CR, Terdiman JP, et al. Revised bethesda guidelines for hereditary nonpolyposis colorectal cancer (lynch syndrome) and microsatellite instability. *J Natl Cancer Inst*. 2004;96(4):261-268.

25. Lindor NM, Rabe K, Petersen GM, et al. Lower cancer incidence in amsterdam-I criteria families without mismatch repair deficiency: Familial colorectal cancer type X. *JAMA*. 2005;293(16):1979-1985.

26. Valle L. Genetic predisposition to colorectal cancer: Where we stand and future perspectives. *World J Gastroenterol*. 2014;20(29):9828-9849.

27. Jasperson KW, Tuohy TM, Neklason DW, Burt RW. Hereditary and familial colon cancer. *Gastroenterology*. 2010;138(6):2044-2058.

28. Groden J, Thliveris A, Samowitz W, et al. Identification and characterization of the familial adenomatous polyposis coli gene. *Cell*. 1991;66(3):589-600.

29. Kinzler KW, Nilbert MC, Su LK, et al. Identification of FAP locus genes from chromosome 5q21. *Science*. 1991;253(5020):661.

30. Brown GR, Hem V, Katz KS, et al. Gene: A gene-centered information resource at NCBI. *Nucleic Acids Res*. 2014;43:D36-D42.

31. Nieuwenhuis MH, Vasen HFA. Correlations between mutation site in APC and phenotype of familial adenomatous polyposis (FAP): A review of the literature. *Crit Rev Oncol Hematol*. 2007;61(2):153-161.

32. Half E, Bercovich D, Rozen P. Familial adenomatous polyposis. *Orphanet J Rare Dis*. 2009;4:22-22.

33. Burt RW, Leppert MF, Slattery ML, et al. Genetic testing and phenotype in a large kindred with attenuated familial adenomatous polyposis. *Gastroenterology*. 2004;127(2):444-451.

34. Sampson JR, Dolwani S, Jones S, et al. Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet*. 2003;362(9377):39-41.

35. Morak M, Laner A, Bacher U, Keiling C, Holinski-Feder E. MUTYH-associated polyposis – variability of the clinical phenotype in patients with biallelic and monoallelic MUTYH mutations and report on novel mutations. *Clin Genet*. 2010;78(4):353-363.

36. Hemminki A, Markie D, Tomlinson I, et al. A serine/threonine kinase gene defective in peutz-jeghers syndrome. *Nature*. 1998;391(6663):184-187.

37. Jenne DE, Reomann H, Nezu J, et al. Peutz-jeghers syndrome is caused by mutations in a novel serine threoninekinase. *Nat Genet*. 1998;18(1):38-43.

38. Peutz JLA. Very remarkable case of familial polyposis of mucous membrane of intestinal tract and nasopharynx accompanied by peculiar pigmentations of skin and mucous membrane. *Nederl Maandschr Geneesk*. 1921;10:134-146.

39. Jeghers H, McKusick VA, Katz KH. Generalized intestinal polyposis and melanin spots of the oral mucosa, lips and digits. *N Engl J Med*. 1949;241(25):993-1005.

40. Tomlinson IP, Houlston RS. Peutz-jeghers syndrome. *J Med Genet*. 1997;34(12):1007.

41. Hearle N, Schumacher V, Menko FH, et al. Frequency and spectrum of cancers in the peutz-jeghers syndrome. *Clin Cancer Res*. 2006;12(10):3209.

42. Houlston R, Bevan S, Williams A, et al. Mutations in DPC4 (SMAD4) cause juvenile polyposis syndrome, but only account for a minority of cases. *Hum Mol Genet*. 1998;7(12):1907-1912.

43. Howe JR, Bair JL, Sayed MG, et al. Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat Genet*. 2001;28(2):184-187.

44. Gardner EJ. A genetic and clinical study of intestinal polyposis, a predisposing factor for carcinoma of the colon and rectum. *Am J Hum Genet*. 1951;3(2):167-176.

45. Brosens LAA, van Hattem A, Hylind LM, et al. Risk of colorectal cancer in juvenile polyposis. *Gut*. 2007;56(7):965-967.

46. Howe JR, Mitros FA, Summers RW. The risk of gastrointestinal carcinoma in familial juvenile polyposis. *Ann Surg Oncol*. 1998;5(8):751-756.

47. Lindblom A, Zhou X, Liu T, Liljegren A, Skoglund J, Djureinovic T. Colorectal cancer as a complex disease: Defining at-risk subjects in the general population - a preventive strategy. *Expert Rev Anticancer Ther*. 2004;4(3):377-385.

48. Roepman P, Schlicker A, Tabernero J, et al. Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition. *Int J Cancer*. 2014;134(3):552-562.

49. Rodriguez-Salas N, Dominguez G, Barderas R, et al. Clinical relevance of colorectal cancer molecular subtypes. *Crit Rev Oncol*. 2017;109:9-19.

50. Budinska E, Popovici V, Tejpar S, et al. Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *J Pathol*. 2013;231(1):63-76.

51. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med*. 2015;21(11):1350-1356

52. Yamagishi H, Kuroda H, Imai Y, Hiraishi H. Molecular pathogenesis of sporadic colorectal cancers. *Chin J Cancer*. 2016;35(1):4.

53. Pino MS, Chung DC. The chromosomal instability pathway in colon cancer. *Gastroenterology*. 2010;138(6):2059-2072.

54. Watanabe T, Kobunai T, Yamamoto Y, et al. Chromosomal instability (CIN) phenotype, CIN high or CIN low, predicts survival for colorectal cancer. *J Clin Oncol*. 2012;30(18):2256-2264.

55. Mouradov D, Domingo E, Gibbs P, et al. Survival in stage II/III colorectal cancer is independently predicted by chromosomal and microsatellite instability, but not by specific driver mutations. *Am J Gastroenterol*. 2013;108(11):1785-1793.

56. Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology*. 2010;138(6):2073-2087.e3.

57. Boland CR, Thibodeau SN, Hamilton SR, et al. A national cancer institute workshop on microsatellite instability for cancer detection and familial predisposition: Development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res*. 1998;58(22):5248.

58. Carethers JM, Chauhan DP, Fink D, et al. Mismatch repair proficiency and in vitro response to 5-fluorouracil. *Gastroenterology*. 1999;117(1):123-131.

59. Popat S, Hubner R, Houlston RS. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol*. 2005;23(3):609-618.

60. Samowitz WS, Albertsen H, Herrick J, et al. Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. *Gastroenterology*. 2005;129(3):837-845.

61. Bae JM, Kim MJ, Kim JH, et al. Differential clinicopathological features in microsatellite instability-positive colorectal cancers depending on CIMP status. *Virchows Archiv*. 2011;459(1):55-63.

62. Barault L, Charon-Barra C, Jooste V, et al. Hypermethylator phenotype in sporadic colon cancer: Study on a population-based series of 582 cases. *Cancer Res*. 2008;68(20):8541.

63. Shen L, Toyota M, Kondo Y, et al. Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proc Natl Acad Sci U S A*. 2007;104(47):18654-18659.

64. Juo YY, Johnston FM, Zhang DY, et al. Prognostic value of CpG island methylator phenotype among colorectal cancer patients: A systematic review and meta-analysis. *Ann Oncol*. 2014;25(12):2314-2327.

65. Bosman FT, Carneiro F, Hruban RH, Theise ND, eds. Chapter 8: Tumors of the colon and rectum. *WHO classification of tumors of the digestive system.* 4th ed. Lyon, France: International Agency for Research on Cancer; 2010.

66. American Cancer Society. Colorectal cancer. http://www.cancer.org/acs/groups/cid/documents/webcontent/003096-pdf.pdf. Updated 2016. Accessed 2017.

67. Symonds DA, Vickery AL. Mucinous carcinoma of the colon and rectum. *Cancer*. 1976;37(4):1891-1900.

68. Odone V, Chang L, Caces J, George SL, Pratt CB. The natural history of colorectal carcinoma in adolescents. *Cancer*. 1982;49(8):1716-1720.

69. Consorti F, Lorenzotti A, Midiri G, Di Paola M. Prognostic significance of mucinous carcinoma of colon and rectum: A prospective case-control study. *J Surg Oncol*. 2000;73(2):70-74.

70. Catalano V, Loupakis F, Graziano F, et al. Mucinous histology predicts for poor response rate and overall survival of patients with colorectal cancer and treated with first-line oxaliplatin- and/or irinotecan-based chemotherapy. *Br J Cancer*. 2009;100(6):881-887.

71. Negri FV, Wotherspoon A, Cunningham D, Norman AR, Chong G, Ross PJ. Mucinous histology predicts for reduced fluorouracil responsiveness and survival in advanced colorectal cancer. *Ann Oncol*. 2005;16(8):1305-1310.

72. Fleming M, Ravula S, Tatishchev SF, Wang HL. Colorectal carcinoma: Pathologic aspects. *J Gastrointest Oncol*. 2012;3(3):153-173.

73. Anthony T, George R, Rodriguez-Bigas M, Petrelli NJ. Primary signet-ring cell carcinoma of the colon and rectum. *Ann Surg Oncol*. 1996;3(4):344-348.

74. Sung CO, Seo JW, Kim K, Do I, Kim SW, Park C. Clinical significance of signet-ring cells in colorectal mucinous adenocarcinoma. *Mod Pathol*. 2008;21(12):1533-1541.

75. Kakar S, Smyrk TC. Signet ring cell carcinoma of the colorectum: Correlations between microsatellite instability, clinicopathologic features and survival. *Mod Pathol*. 2004;18(2):244-249.

76. Thota R, Fang X, Subbiah S. Clinicopathological features and survival outcomes of primary signet ring cell and mucinous adenocarcinoma of colon: Retrospective analysis of VACCR database. *J Gastrointest Oncol*. 2013;5(1):18-24.

77. Thirunavukarasu P, Sathaiah M, Singla S, et al. Medullary carcinoma of the large intestine: A population based analysis. *Int J Oncol*. 2010;37(4):901-907.

78. Lanza G, Gafà R, Matteuzzi M, Santini A. Medullary-type poorly differentiated adenocarcinoma of the large bowel: A distinct clinicopathologic entity characterized by microsatellite instability and improved survival. *JCO*. 1999;17(8):2429-2429.

79. Erstad DJ, Tumusiime G, Cusack JCJ. Prognostic and predictive biomarkers in colorectal cancer: Implications for the clinical surgeon. *Ann Surg Oncol*. 2015;22(11):3433-3450.

80. Edge S, Byrd DR, Compton CC, Fritz AG, Greene FL, Trotti A, eds. *AJCC (american joint committee on cancer) Cancer Staging Handbook* . 7th ed. New York: Springer; 2010.

81. Liang P, Nakada I, Hong J, et al. Prognostic significance of immunohistochemically detected blood and lymphatic vessel invasion in colorectal carcinoma: Its impact on prognosis. *Ann Surg Oncol*. 2007;14(2):470-477.

82. Ueno H, Murphy J, Jass JR, Mochizuki H, Talbot IC. Tumour `budding' as an index to estimate the potential of aggressiveness in rectal cancer. *Histopathology*. 2002;40(2):127-132.

83. Ohtsuki K, Koyama F, Tamura T, et al. Prognostic value of immunohistochemical analysis of tumor budding in colorectal carcinoma. *Anticancer Res*. 2008;28(3B):1831-1836.

84. Okuyama T, Nakamura T, Yamaguchi M. Budding is useful to select high-risk patients in stage II well-differentiated or moderately differentiated colon adenocarcinoma. *Dis Colon Rectum*. 2003;46(10):1400-1406.

85. Canadian Cancer Society. Grading colorectal cancer. http://www.cancer.ca/en/cancer-information/cancer-type/colorectal/grading/?region=on. Updated 2016. Accessed 2017.

86. Compton CC. Colorectal carcinoma: Diagnostic, prognostic, and molecular features. *Mod Pathol*. 2003;16(4):376-388.

87. Compton CC, Fielding LP, Burgart LJ, et al. Prognostic factors in colorectal cancer. *Arch Pathol Lab Med*. 2000;124(7):979-994.

88. Roncucci L, Fante R, Losi L, et al. Survival for colon and rectal cancer in a population-based cancer registry. *Eur J Cancer*. 1996;32(2):295-302.

89. Enblad P, Adami H, Bergström R, Glimelius B, Krusemo U, Påhlman L. Improved survival of patients with cancers of the colon and rectum? *J Natl Cancer Inst*. 1988;80(8):586-591.

90. Kapiteijn E, Marijnen CAM, Colenbrander AC, et al. Local recurrence in patients with rectal cancer diagnosed between 1988 and 1992: A population-based study in the west netherlands. *Eur J Surg Oncol*. 1998;24(6):528-535.

91. Lee Y, Lee Y, Chuang J, Lee J. Differences in survival between colon and rectal cancer from SEER data. *PLoS ONE*. 2013;8(11):e78709.

92. Joern F, Gunter H, Thomas J, et al. Outcome for stage II and III rectal and colon cancer equally good after treatment improvement over three decades. *Int J Colorectal Dis*. 2015;30(6):797-806.

93. van dS, Bastiaannet E, Mesker WE, et al. Differences between colon and rectal cancer in complications, short-term survival and recurrences. *Int J Colorectal Dis*. 2016;31(10):1683-1691.

94. McKay A, Donaleshen J, Helewa RM, et al. Does young age influence the prognosis of colorectal cancer: A population-based analysis. *World J Surg Oncol*. 2014;12:370.

95. Li J, Wang Z, Yuan X, Xu L, Tong J. The prognostic significance of age in operated and non-operated colorectal cancer. *BMC Cancer*. 2015;15:83.

96. Majek O, Gondos A, Jansen L, et al. Sex differences in colorectal cancer survival: Population-based analysis of 164,996 colorectal cancer patients in germany. *PLoS ONE*. 2013;8(7):e68077.

97. Wichmann MW, Müller C, Hornung HM, Lau-Werner U, Schildberg FW. Gender differences in long-term survival of patients with colorectal cancer. *Br J Surg*. 2001;88(8):1092-1098.

98. McArdle CS, McMillan DC, Hole DJ. Male gender adversely affects survival following surgery for colorectal cancer. *Br J Surg*. 2003;90(6):711-715.

99. Walter V, Jansen L, Hoffmeister M, Brenner H. Smoking and survival of colorectal cancer patients: Systematic review and meta-analysis. *Ann Oncol*. 2014;25(8):1517-1525.

100. Meyerhardt JA, Heseltine D, Niedzwiecki D, et al. Impact of physical activity on cancer recurrence and survival in patients with stage III colon cancer: Findings from CALGB 89803. *JCO*. 2006;24(22):3535-3541.

101. Meyerhardt JA, Giovannucci EL, Holmes MD, et al. Physical activity and survival after colorectal cancer diagnosis. *JCO*. 2006;24(22):3527-3534.

102. The Cancer Genome AN. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330-337.

103. Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. *Cell Res*. 2002;12(1):9-18.

104. Tejpar S, Bertagnolli M, Bosman F, et al. Prognostic and predictive biomarkers in resected colon cancer: Current status and future perspectives for integrating genomics into biomarker discovery. *Oncologist*. 2010;15(4):390-404.

105. Allegra CJ, Jessup JM, Somerfield MR, et al. American society of clinical oncology provisional clinical opinion: Testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to Anti–Epidermal growth factor receptor monoclonal antibody therapy. *JCO*. 2009;27(12):2091-2096.

106. Davies H, Bignell GR, Cox C, et al. Mutations of the BRAF gene in human cancer. *Nature*. 2002;417(6892):949-954.

107. Fariña-Sarasqueta A, van Lijnschoten G, Moerland E, et al. The BRAF V600E mutation is an independent prognostic factor for survival in stage II and stage III colon cancer patients. *Ann Oncol*. 2010;21(12):2396-2402.

108. Phipps AI, Buchanan DD, Makar KW, et al. KRAS-mutation status in relation to colorectal cancer survival: The joint impact of correlated tumour markers. *Br J Cancer*. 2013;108(8):1757-1764.

109. Kadowaki S, Kakuta M, Takahashi S, et al. Prognostic value of KRAS and BRAF mutations in curatively resected colorectal cancer. *World J Gastroenterol*. 2014;21(4):1275-1283.

110. Ting W, Chen L, Huang L, et al. Impact of interleukin-10 gene polymorphisms on survival in patients with colorectal cancer. *J Korean Med Sci*. 2013;28(9):1302-1306.

111. Pardini B, Bermejo JL, Naccarati A, et al. Inherited variability in a master regulator polymorphism (rs4846126) associates with survival in 5-FU treated colorectal cancer patients. *Mutat Res*. 2014;766–767:7-13.

112. Dai J, Wan S, Zhou F, et al. Genetic polymorphism in a VEGF-independent angiogenesis gene ANGPT1 and overall survival of colorectal cancer patients after surgical resection. *PLoS ONE*. 2012;7(4):e34758.

113. Inoue Y, Hazama S, Iwamoto S, et al. FcÎ³R and EGFR polymorphisms as predictive markers of cetuximab efficacy in metastatic colorectal cancer. *Mol Diagn Ther*. 2014;18(5):541-548.

114. Jaka A, Gutiérrez-Rivera A, Ormaechea N, et al. Association between EGFR gene polymorphisms, skin rash and response to anti-EGFR therapy in metastatic colorectal cancer patients. *Exp Dermatol*. 2014;23(10):751-753.

115. Etienne M, Formento J, Chazal M, et al. Methylenetetrahydrofolate reductase gene polymorphisms and response to fluorouracil-based treatment in advanced colorectal cancer patients. *Pharmacogenetics*. 2004;14(12):785-792.

116. Li A, Meyre D. Challenges in reproducibility of genetic association studies: Lessons learned from the obesity field. *Int J Obes*. 2013;37(4):559-567.

117. Griffiths AJF. Chapter 1: Genetics and the Organism. *An Introduction to Genetic Analysis.* 7th ed. New York: New York : W.H. Freeman; 2000.

118. Genetics Home Reference. Help me understand genetics - genomic research. https://ghr.nlm.nih.gov/primer/genomicresearch/snp. Updated 2017.

119. Strachan T. Chapter 13: Human Genetic Variability and its Consequences. *Human molecular genetics.* 4th ed. New York: New York : Garland Science; 2011.

120. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74.

121. Wollstein A, Stephan W. Inferring positive selection in humans from genomic data. *Investigative Genetics*. 2015;6(1):5.

122. Vasseur E, Quintana-Murci L. The impact of natural selection on health and disease: Uses of the population genetics approach in humans. *Evol Appl*. 2013;6(4):596-607.

123. Keinan A, Mullikin JC, Patterson N, Reich D. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in east Asians than in Europeans. *Nat Genet*. 2007;39(10):1251-1255.

124. Rosenberg NA. A population-genetic perspective on the similarities and differences among worldwide human populations. *Hum Biol*. 2011;83(6):659-684.

125. Morrison JA, Gruppo R, Glueck CJ, et al. Population-specific alleles: The polymorphism (k121q) of the human glycoprotein PC-1 gene is strongly associated with race but not with insulin resistance in black and white children. *Metabolism*. 2004;53(4):465-468.

126. Turchin MC, Chiang CWK, Palmer CD, et al. Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat Genet.* 2012;44:1015.

127. Mattei J, Parnell LD, Lai C, et al. Disparities in allele frequencies and population differentiation for 101 disease-associated single nucleotide polymorphisms between Puerto Ricans and non-Hispanic whites. *BMC Genetics*. 2009;10(1):45.

128. Li G, Xu R, Cao Y, Xie X, Zheng Z. Interleukin-21 polymorphism affects gene expression and is associated with risk of ischemic stroke. *Inflammation*. 2014;37(6):2030-2039.

129. Stiers KM, Graham AC, Kain BN, Beamer LJ. Asp263 missense variants perturb the active site of human phosphoglucomutase 1. *The FEBS Journal*. 2017;284(6):937-947.

130. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. *Nature*. 2006;444(7118):444-454.

131. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: Changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet*. 2006;15:R57-R66.

132. Falchi M, El-Sayed Moustafa JS, Takousis P, et al. Low copy number of the salivary amylase gene predisposes to obesity. *Nat Genet*. 2014;46:492-497.

133. Perry GH, Dominy NJ, Claw KG, et al. Diet and the evolution of human amylase gene copy number variation. *Nat Genet*. 2007;39(10):1256-1260.

134. Gonzalez E, Kulkarni H, Bolivar H, et al. The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*. 2005;307(5714):1434.

135. Yang Y, Chung EK, Wu YL, et al. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor

against SLE susceptibility in European Americans. *Am J Hum Genet*. 2007;80(6):1037-1054.

136. Hollox EJ, Huffmeier U, Zeeuwen PLJM, et al. Psoriasis is associated with increased β-defensin genomic copy number. *Nat Genet*. 2007;40:23-25.

137. Marshall CR, Scherer SW. Detection and characterization of copy number variation in autism spectrum disorder. In: Feuk L, ed. *Genomic structural variants: Methods and protocols.* New York, NY: Springer New York; 2012:115-135.

138. The Human Genome Structural Variation,Working Group, Eichler EE, Nickerson DA, et al. Completing the map of human genetic variation: A plan to identify and integrate normal structural variation into the human genome sequence. *Nature*. 2007;447(7141):161-165.

139. The International HapMap Consortium. A haplotype map of the human genome. *Nature*. 2005;437(7063):1299-1320.

140. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet*. 2001;17(9):502-510.

141. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev*. 2010;11:415.

142. Fisher R. *The genetical theory of natural selection.* Oxford: Oxford Univ. Press; 1930.

143. Feldman MW, Lewontin RC. The heritability hang-up. *Science*. 1975;190(4220):1163.

144. Gray IC, Campbell DA, Spurr NK. Single nucleotide polymorphisms as tools in human genetics. *Hum Mol Genet*. 2000;9(16):2403-2408.

145. Weatherall D, Hofman K, Rodgers G, Ruffin J, Hrynkow S. A case for developing north-south partnerships for research in sickle cell disease. *Blood*. 2005;105(3):921-923.

146. O'Sullivan BP, Freedman SD. Cystic fibrosis. *The Lancet*. 2009;373(9678):1891-1904.

147. Chong J, Buckingham K, Jhangiani S, et al. The genetic basis of mendelian phenotypes: Discoveries, challenges, and opportunities. *Am J Hum Genet*. 2015;97(2):199-215.

148. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science*. 1994;265(5181):2037-2048.

149. Foulkes AS. Genetic association studies. In: Foulkes AS, ed. *Applied statistical genetics with R: For population-based association studies.* New York, NY: Springer New York; 2009:1-27.

150. Lewis CM, Knight J. Introduction to genetic association studies. *Cold Spring Harb Protoc*. 2012;2012(3):pdb.top068163.

151. Thomas DC. Chapter 11: Gene Characterization. *Statistical Methods in Genetic Epidemiology*. Cary: Oxford University Press; 2004.

152. Foulkes AS. Elementary statistical principles. In: Foulkes AS, ed. *Applied statistical genetics with R: For population-based association studies.* New York, NY: Springer New York; 2009:29-63.

153. Wilke RA, Mareedu RK, Moore JH. The pathway less traveled: Moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. *Curr Pharmacogenomics Person Med*. 2008;6(3):150-159.

154. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860-921.

155. Ventimiglia G, Petralia S. Recent advances in DNA microarray technology: An overview on production strategies and detection methods. *Bionanoscience*. 2013;3(4):428-450.

156. Cooper GM, Johnson JA, Langaee TY, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. *Blood*. 2008;112(4):1022-1027.

157. Xie P, Kranzler HR, Yang C, Zhao H, Farrer LA, Gelernter J. Genome-wide association study identifies new susceptibility loci for posttraumatic stress disorder. *Biol Psychiatry*. 2013;74(9):10.1016

158. Athanasiu L, Smorr LH, Tesli M, et al. Genome-wide association study identifies common variants associated with pharmacokinetics of psychotropic drugs. *J Psychopharmacol*. 2015;29(8):884-891.

159. Bush WS, Moore JH. Chapter 11: Genome-wide association studies. *PLoS Computat Biol*. 2012;8(12):e1002822.

160. Montana G. Statistical methods in genetics. *Brief Bioinform*. 2006;7(3):297-308.

161. Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science*. 1996;273(5281):1516.

162. Lewis CM. Genetic association studies: Design, analysis and interpretation. *Brief Bioinform*. 2002;3(2):146-153.

163. Lettre G, Lange C, Hirschhorn JN. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiol*. 2007;31(4):358-362.

164. Bagos PG. Genetic model selection in genome-wide association studies: Robust methods and the use of meta-analysis. *Stat Appl Genet Mol Biol*. 2013;12:285.

165. Zintzaras E, Lau J. Synthesis of genetic association studies for pertinent gene–disease associations requires appropriate methodological and statistical approaches. *J Clin Epidemiol*. 2008;61(7):634-645.

166. Thomas DC. Chapter 4: Basic Epidemiologic and Statistic Principles. *Statistical Methods in Genetic Epidemiology.* Cary: Oxford University Press; 2004.

167. Hosmer DW, Lemeshow S, Sturdivant RX. Chapter 1: Introduction to the Logistic Regression Model. *Applied logistic regression.* New York: John Wiley & Sons, Incorporated; 2013.

168. Williams GM, Ware R. Modelling binary outcomes. In: Williams GM, eds. *Methods of clinical epidemiology*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013:141-163.

169. Jewell NP. Chapter 12: Regression Models Relating Exposure to Disease. *Statistics for epidemiology*. Chapman & Hall/CRC; 2004.

170. Hosmer DW, Lemeshow S, Sturdivant RX. Chapter 2: The Multiple Logistic Regression Model. *Applied logistic regression.* New York: John Wiley & Sons, Incorporated; 2013.

171. Jewell NP. Chapter 13: Estimation of Logistic Regression Model Parameters. *Statistics for epidemiology.* Chapman & Hall/CRC; 2004.

172. Core Team R. R: A language and environment for statistical computing. R foundation for statistical computing. 2013.

173. Zhou X, Obuchowski NA, McClish DK. Chapter 2. Measures of diagnostic accuracy. In: *Statistical methods in diagnostic medicine.* 2nd ed. Hoboken: Wiley; 2011:13-57.

174. Zhou X, Obuchowski NA, McClish DK. Chapter 8. Regression analysis for independent ROC data. In: *Statistical methods in diagnostic medicine.* 2nd ed. Hoboken: Wiley; 2011:263-297.

175. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *J Natl Cancer Inst*. 2003;95(7):511-515.

176. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin Chem*. 1993;39(4):561.

177. Lasko TA, Bhagwat JG, Zou KH, Ohno-Machado L. The use of receiver operating characteristic curves in biomedical informatics. *J Biomed Inform*. 2005;38(5):404-415.

178. Kumar R, Indrayan A. Receiver operating characteristic (ROC) curve for medical researchers. *Indian Pediatr*. 2011;48(4):277-287.

179. Li J, Fine JP. ROC analysis with multiple classes and multiple tests: Methodology and its application in microarray studies. *Biostatistics*. 2008;9(3):566-576.

180. Zheng Y, Cai T, Feng Z. Application of the time-dependent ROC curves for prognostic accuracy with multiple biomarkers. *Biometrics*. 2006;62(1):279-287.

181. Zou KH, O'Malley AJ, Mauri L. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007;115(5):654.

182. Musial J, Swadzba J, Motyl A, Iwaniec T. Clinical significance of antiphospholipid protein antibodies. receiver operating characteristics plot analysis. *J Rheumatol*. 2003;30(4):723.

183. Cheung R, Altschuler MD, D’Amico AV, Malkowicz SB, Wein AJ, Whittington R. Using the receiver operating characteristic curve to select pretreatment and pathologic predictors for early and late postprostatectomy PSA failure. *Urology*. 2001;58(3):400-405.

184. Gail MH. Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *J Natl Cancer Inst*. 2008;100(14):1037-1041.

185. Pepe MS, Janes HE. Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *J Natl Cancer Inst*. 2008;100(14):978-979.

186. Walter SD, Sinuff T. Studies reporting ROC curves of diagnostic and prediction data can be incorporated into meta-analyses using corresponding odds ratios. *J Clin Epidemiol*. 2007;60(5):530-534.

187. Robin X, Turck N, Hainard A, et al. pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77.

188. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311-321.

189. Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet*. 2001;69(1):124-137.

190. The UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature*. 2015;526(7571):82-90.

191. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*. 2004;305(5685):869.

192. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science (New York, N.Y.)*. 2009;324(5925):387-389.

193. Steinthorsdottir V, Thorleifsson G, Sulem P, et al. Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. *Nat Genet*. 2014;46:294.

194. Bonnefond A, Clément N, Fawcett K, et al. Rare MTNR1B variants impairing melatonin receptor 1B function contribute to type 2 diabetes. *Nat Genet*. 2012;44(3):297-301.

195. Ji W, Foo JN, O'Roak B,J., et al. Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat Genet*. 2008;40:592.

196. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*. 2011;43(11):1066-1073.

197. Holm H, Gudbjartsson DF, Sulem P, et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet*. 2011;43(4):316-320.

198. Genovese G, Fromer M, Stahl EA, et al. Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat Neurosci*. 2016;19(11):1433-1441.

199. Lee S, Abecasis G, Boehnke M, Lin X. Rare-variant association analysis: Study designs and statistical tests. *Am J Hum Genet*. 2014;95(1):5-23.

200. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5(2):e1000384. is it PLoS or PLOS?

201. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat Res*. 2007;615(1–2):28-56.

202. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol*. 2011;35(7):606-619.

203. Neale BM, Rivas MA, Voight BF, et al. Testing for an unusual distribution of rare variants. *PLoS Genet*. 2011;7(3):e1001322.

204. Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93.

205. Ionita-Laza I, Lee S, Makarov V, Buxbaum J, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. *Am J Hum Genet*. 2013;92(6):841-853.

206. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012;13(4):762-775.

207. Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genet Epidemiol*. 2013;37(1):110-121.

208. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. *Genetics*. 2014;197(4):1081.

209. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidemiol*. 2013;37(4):334-344.

210. Lee S, Emond MJ, Bamshad MJ, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012;91(2):224-237.

211. Lee S, with contributions from Miropolsky, L. and Wu, M. SKAT: SNP-set (sequence) kernel association test. R package version 1.2.1. https://CRAN.R-project.org/package=SKAT. Updated 2016.

212. Williams GM, Ware R. Modelling time-to-event data. In: Doi SAR, Williams GM, eds. *Methods of clinical epidemiology*. Berlin, Heidelberg: Springer Berlin Heidelberg; 2013:165-183.

213. Lee ET, Wang JW. *Statistical methods for survival data analysis.* Somerset: Wiley; 2013.

214. Sy JP, Taylor JMG. Estimation in a cox proportional hazards cure model. *Biometrics*. 2000;56(1):227-236.

215. Paoletti X, Asselain B. Survival analysis in clinical trials: Old tools or new techniques. *Surg Oncol*. 2010;19(2):55-58.

216. Farewell VT. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*. 1982;38(4):1041-1046.

217. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc*. 1958;53(282):457-481.

218. Therneau T. A package for survival analysis in S. version 2.38. http://CRAN.R-project.org/package=survival. Updated 2015.

219. Cox DR. Regression models and life-tables. *J R Stat Soc Ser B Methodol*. 1972;34(2):187-220.

220. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*. 1994;81(3):515-526.

221. Bejan-Angoulvant T, Bouvier A, Bossard N, et al. Hazard regression model and cure rate model in colon cancer relative survival trends: Are they telling the same story? *Eur J Epidemiol*. 2008;23(4):251-259.

222. Yilmaz YE, Lawless JF, Andrulis IL, Bull SB. Insights from mixture cure modeling of molecular markers for prognosis in breast cancer. *J Clin Oncol*. 2013;31(16):2047-2054.

223. Forse C, Yilmaz Y, Pinnaduwage D, et al. Elevated expression of podocalyxin is associated with lymphatic invasion, basal-like phenotype, and clinical outcome in axillary lymph node-negative breast cancer. *Breast Cancer Res Treat*. 2013;137(3):709-719.

224. Lambert PC, Dickman PW, Weston CL, Thompson JR. Estimating the cure fraction in population-based cancer studies by using finite mixture models. *J R Stat Soc Ser C Appl Stat*. 2010;59(1):35-55.

225. Tsodikov AD, Ibrahim JG, Jakovlev AY. Estimating cure rates from survival data. *J Am Stat Assoc*. 2003;98(464):1063-1078.

226. Cox DR. Summary comments. *Surg Oncol*. 2010;19(2):61.

227. Lawless J. *Statistical models and methods for lifetime data.* 2nd, Wiley series in probability and statistics ed. Hoboken, N.J.: Wiley-Interscience; 2003.

228. Brown M, Tsodikov A, Bauer KR, Parise CA, Caggiano V. The role of human epidermal growth factor receptor 2 in the survival of women with estrogen and progesterone receptor-negative, invasive breast cancer: The California cancer registry, 1999–2004. *Cancer*. 2008;112(4):737-747.

229. Broët P, Kuznetsov VA, Bergh J, Liu ET, Miller LD. Identifying gene expression changes in breast cancer that distinguish early and late relapse among uncured patients. *Bioinformatics*. 2006;22(12):1477-1485.

230. Anthony YCK, Chen-Hsin Chen. A mixture model combining logistic regression with proportional hazards regression. *Biometrika*. 1992;79(3):531-541.

231. Peng Y, Dear KB, Denham J. A generalized F mixture model for cure rate estimation. *Stat Med*. 1998;17(8):813-830.

232. Peng Y, Keith BGD. A nonparametric mixture model for cure rate estimation. *Biometrics*. 2000;56(1):237-243.

233. Zhao Y, Lee AH, Yau KKW, Burke V, McLachlan GJ. A score test for assessing the cured proportion in the long-term survivor mixture model. *Stat Med*. 2009;28(27):3454-3466.

234. Lawless JF, Yilmaz YE. Semiparametric estimation in copula models for bivariate sequential survival times. *Biom J*. 2011;53(5):779-796.

235. Berkson J, Gage RP. Survival curve for cancer patients following treatment. *J Am Stat Assoc*. 1952;47(259):501-515.

236. Farewell VT. Mixture models in survival analysis: Are they worth the risk? *Can J Stat*. 1986;14(3):257-262.

237. Jeremy MGT. Semi-parametric estimation in failure time mixture models. *Biometrics*. 1995;51(3):899-907.

238. Taylor J. Semi-parametric estimation in failure time mixture models. *Biometrics*. 1995;51(3):899-907.

239. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Parzen E, Tanabe K, Kitagawa G, eds. *Selected papers of hirotugu akaike.* New York, NY: Springer New York; 1998:199-213.

240. Neyman J, Pearson ES. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London.Series A, Containing Papers of a Mathematical or Physical Character*. 1933;231:289-337.

241. Okudaira K, Kakar S, Cun L, et al. MUC2 gene promoter methylation in mucinous and non-mucinous colorectal cancer tissues. *Int J Oncol*. 2010;36(4):765-775.

242. Hanski C. Is mucinous carcinoma of the colorectum a distinct genetic entity? *Br J Cancer*. 1995;72(6):1350-1356.

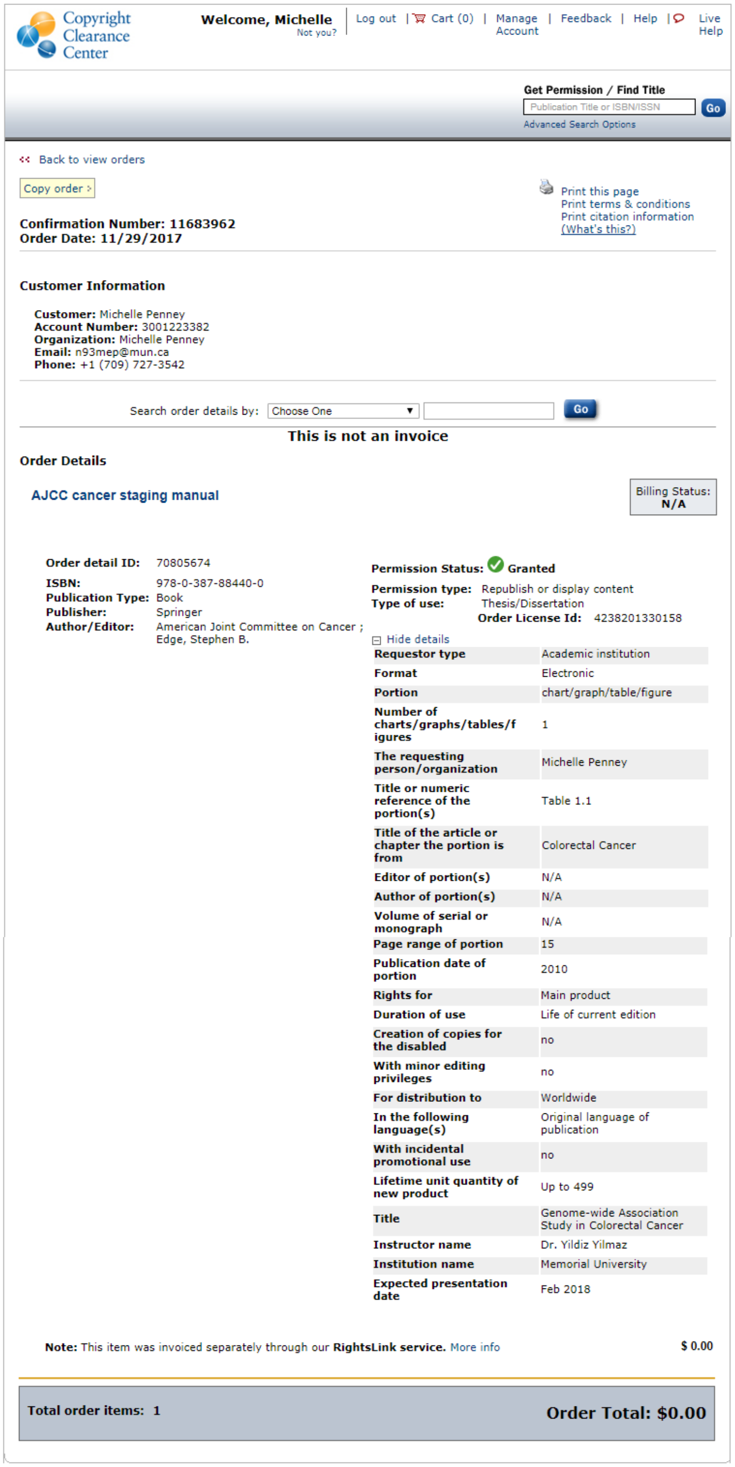
243. Kim DH, Kim JW, Cho JH, et al. Expression of mucin core proteins, trefoil factors, APC and p21 in subsets of colorectal polyps and cancers suggests a distinct pathway of pathogenesis of mucinous carcinoma of the colorectum. *Int J Oncol*. 2005;27:957-964.

244. Monteiro J, Fodde R. Cancer stemness and metastasis: Therapeutic consequences and perspectives. *Eur J Cancer*. 2010;46(7):1198-1203.

245. Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. *Science*. 2011;331(6024):1559.

APPENDICES

Appendix A: Copyright approval for the use of the table from AJCC


Copyright Clearance Center
Welcome, Michelle
Not you?
Log out | Cart (0) | Manage Account | Feedback | Help | Live Help
Get Permission / Find Title
Publication Title or ISBN/ISSN
Go
Advanced Search Options
Back to view orders
Copy order
Print this page
Print terms & conditions
Print citation information
(What's this?)
Search order details by: Choose One
Go

Confirmation Number: 11683962
Order Date: 11/29/2017

Customer Information

Customer: Michelle Penney
Account Number: 3001223382
Organization: Michelle Penney
Email: n93mep@mun.ca
Phone: +1 (709) 727-3542

This is not an invoice

Order Details

AJCC cancer staging manual

Billing Status: **N/A**

Order detail ID: 70805674
ISBN: 978-0-387-88440-0
Publication Type: Book
Publisher: Springer
Author/Editor: American Joint Committee on Cancer ; Edge, Stephen B.

Permission Status: ✓ **Granted**
Permission type: Republish or display content
Type of use: Thesis/Dissertation
Order License Id: 4238201330158

⊟ Hide details

Requestor type	Academic institution
Format	Electronic
Portion	chart/graph/table/figure
Number of charts/graphs/tables/figures	1
The requesting person/organization	Michelle Penney
Title or numeric reference of the portion(s)	Table 1.1
Title of the article or chapter the portion is from	Colorectal Cancer
Editor of portion(s)	N/A
Author of portion(s)	N/A
Volume of serial or monograph	N/A
Page range of portion	15
Publication date of portion	2010
Rights for	Main product
Duration of use	Life of current edition
Creation of copies for the disabled	no
With minor editing privileges	no
For distribution to	Worldwide
In the following language(s)	Original language of publication
With incidental promotional use	no
Lifetime unit quantity of new product	Up to 499
Title	Genome-wide Association Study in Colorectal Cancer
Instructor name	Dr. Yildiz Yilmaz
Institution name	Memorial University
Expected presentation date	Feb 2018

Note: This item was invoiced separately through our **RightsLink service.** More info **$ 0.00**

Total order items: 1 **Order Total: $0.00**

Appendix B: Supplementary data for Penney *et al.*, 2018: "Associations of single nucleotide polymorphisms with mucinous colorectal cancer: genome-wide common variant and gene-based rare variant analyses"

Supplementary Table S1. Top ten most significant common SNPs identified based on the univariable analyses and the subsequent multivariable analyses under the additive genetic models.

	Univariable		**Multivariable***	
SNP ID (a vs. b)	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
rs11159673 (AG + 2*AA vs. GG)	3.096 (1.879-5.073)	7.11E-06	3.016 (1.766-5.138)	4.53E-05
rs7314811 (CT + 2*CC vs. TT)	2.387 (1.611-3.567)	1.67E-05	2.129 (1.409-3.242)	3.59E-04
rs4843335 (AG + 2*AA vs. GG)	3.818 (2.040-7.037)	1.97E-05	4.160 (2.130-8.107)	2.64E-05
rs10511330 (CT + 2*CC vs. TT)	3.013 (1.788-5.031)	2.59E-05	3.908 (2.211-6.945)	2.65E-06
rs12915222 (CT + 2*TT vs. CC)	2.476 (1.623-3.801)	2.76E-05	2.664 (1.685-4.260)	3.19E-05
rs12956191 (GA + 2*GG vs. AA)	2.364 (1.580-3.549)	2.86E-05	2.397 (1.555-3.726)	8.18E-05
rs11648965 (GA + 2*GG vs. AA)	3.033 (1.786-5.104)	3.04E-05	3.250 (1.867-5.671)	2.82E-05
rs16822593 (AG + 2*AA vs. GG)	2.977 (1.767-4.968)	3.11E-05	3.899 (2.205-6.930)	2.78E-06
rs205536 (CT + 2*CC vs. TT)	2.494 (1.636-3.885)	3.28E-05	2.681 (1.717-4.294)	2.34E-05
rs2384298 (CT + 2*CC vs. TT)	2.332 (1.563-3.489)	3.42E-05	2.406 (1.576-3.702)	5.14E-05

CI: confidence interval, OR: odds ratio.
OR is the ratio of the odds of having mucinous tumors for one minor allele increase.
*Multivariable logistic regression model adjusting for the selected baseline characteristics listed in Supplementary Table S5.

Supplementary Table S2. Top ten most significant common SNPs identified based on the univariable analyses and the subsequent multivariable analysis under the dominant genetic models.

	Univariable		**Multivariable***	
SNP ID (a vs. b)	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
rs716897 (CC + CT vs. TT)	0.268 (0.150-0.469)	5.33E-06	0.262 (0.143-0.473)	1.12E-05
rs10511330 (CC + CT vs. TT)	3.771 (2.059-6.809)	1.24E-05	4.851 (2.544-9.232)	1.40E-06
rs11968293 (CC + CA vs. AA)	0.285 (0.161-0.501)	1.27E-05	0.264 (0.143-0.481)	1.48E-05
rs17712784 (AA + AG vs. GG)	3.469 (1.972-6.120)	1.54E-05	3.299 (1.798-6.064)	1.12E-04
rs13019215 (TT + TC vs. CC)	0.267 (0.144-0.479)	1.56E-05	0.232 (0.119-0.432)	8.20E-06
rs16822593 (AA + AG vs. GG)	3.704 (2.024-6.681)	1.59E-05	4.834 (2.534-9.202)	1.50E-06
rs12471607 (TT + TC vs. CC)	0.268 (0.144-0.480)	1.65E-05	0.233 (0.119-0.433)	8.42E-06
rs4843335 (AA + AG vs. GG)	4.108 (2.108-7.794)	2.06E-05	4.672 (2.298-9.344)	1.48E-05
rs11216624 (AA + AG vs. GG)	3.565 (1.933-6.460)	3.34E-05	3.074 (1.604-5.787)	5.67E-04
rs9809129 (AA + AG vs. GG)	0.231 (0.107-0.449)	4.78E-05	0.226 (0.103-0.451)	6.52E-05

CI: confidence interval, OR: odds ratio.
OR compares the odds of having mucinous tumors in subgroup a to the odds of having mucinous tumors in subgroup b. *Multivariable logistic regression model adjusting for the selected baseline characteristics listed in Supplementary Table S5.

Supplementary Table S3. Top ten most significant common SNPs identified based on the univariable analyses and the subsequent multivariable analyses under the recessive genetic models.

	Univariable		**Multivariable***	
SNP ID (a vs. b)	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
rs9481067 (GG vs. AG + AA)	4.171 (2.332-7.431)	1.24E-06	4.747 (2.527-8.948)	1.24E-06
rs4837345 (TT vs. TC + CC)	4.721 (2.403-9.052)	4.00E-06	4.563 (2.242-9.107)	1.97E-05
kgp10457679 (CC vs. CT + TT)	4.721 (2.403-9.052)	4.00E-06	4.563 (2.242-9.107)	1.97E-05
kgp4136779 (TT vs. TC + CC)	4.721 (2.403-9.052)	4.00E-06	4.563 (2.242-9.107)	1.97E-05
rs1075650 (GG vs. AG + AA)	4.721 (2.403-9.052)	4.00E-06	4.563 (2.242-9.107)	1.97E-05
rs7314811 (CC vs. CT + TT)	4.586 (2.338-8.772)	5.66E-06	3.654 (1.746-7.425)	4.15E-04
rs6596805 (GG vs. AG + AA)	3.862 (2.139-6.908)	5.72E-06	3.683 (1.957-6.887)	4.51E-05
rs11047047 (GG vs. AG + AA)	3.734 (2.104-6.613)	5.92E-06	3.505 (1.908-6.426)	4.79E-05
rs1661281 (TT vs. TC + CC)	5.284 (2.518-10.764)	6.12E-06	5.403 (2.442-11.704)	2.17E-05
rs919001 (AA vs. AG + GG)	3.913 (2.134-7.077)	7.46E-06	3.151 (1.636-5.953)	4.66E-04

CI: confidence interval, OR: odds ratio.
OR compares the odds of having mucinous tumors in subgroup a to the odds of having mucinous tumors in subgroup b.
*Multivariable logistic regression model adjusting for the selected baseline characteristics listed in Supplementary Table S5.

Supplementary Table S4. Top ten most significant common SNPs identified under the univariable analyses and the subsequent multivariable analyses under the co-dominant genetic models.

	Univariable		**Multivariable***	
SNP ID (a vs. b)	**OR (95% CI)**	**p-value**	**OR (95% CI)**	**p-value**
rs7314811 (CC vs. TT)	5.974 (2.803-12.805)	3.48E-06	4.788 (2.109-10.853)	1.63E-04
rs16907305 (AA vs. GG)	5.550 (2.611-11.857)	7.91E-06	4.505 (1.994-10.162)	2.66E-04
rs11216624 (AG vs. GG)	3.872 (2.092-7.050)	1.15E-05	3.326 (1.727-6.302)	2.57E-04
rs17712784 (AG vs. GG)	3.520 (1.988-6.243)	1.50E-05	3.304 (1.788-6.106)	1.28E-04
rs6573132 (AG vs. GG)	4.814 (2.308-9.722)	1.62E-05	5.183 (2.382-11.030)	2.27E-05
rs8019850 (TC vs. CC)	4.802 (2.302-9.699)	1.67E-05	5.011 (2.310-10.619)	3.07E-05
rs17093005 (TG vs. GG)	4.802 (2.302-9.699)	1.67E-05	5.098 (2.345-10.828)	2.66E-05
rs11656626 (GG vs. AA)	7.194 (2.866-17.642)	1.72E-05	7.156 (2.675-18.933)	6.94E-05
rs1189903 (AC vs. CC)	4.759 (2.284-9.589)	1.78E-05	4.952 (2.283-10.488)	3.47E-05
rs4779810 (TT vs. CC)	5.965 (2.608-13.499)	1.79E-05	4.286 (1.731-10.343)	1.30E-03

CI: confidence interval, OR: odds ratio.
OR compares the odds of having mucinous tumors in subgroup a to the odds of having mucinous tumors in subgroup b.
*Multivariable logistic regression model adjusting for the selected baseline characteristics listed in Supplementary Table S5.

Supplementary Table S5. Baseline characteristics selected through a stepwise variable selection method under the multivariable model.

Characteristics		**OR (95% CI)**	**P-value**
Age	≤60		
	60-65	2.29 (1.08-4.81)	0.018
	>65	1.19 (0.60-2.37)	0.611
Sex	Female		
	Male	0.58 (0.32-1.04)	0.067
Location	Colon		
	Rectum	0.45 (0.21-0.90)	0.031
Stage	I		
	II	4.41 (1.48-18.98)	0.018
	III	3.65 (1.18-16.02)	0.044
	IV	4.57 (1.18-22.43)	0.036
Grade	Well/moderately diff.		
	Poorly diff.	1.90 (0.70-4.54)	0.169

CI: confidence interval, diff.: differentiated.
OR: odds ratio (compares the odds of having mucinous tumors with the corresponding factor level to the odds of having mucinous tumors with the reference factor level).

Supplementary Table S6. AIC estimates under the multivariable models of common SNPs identified in the univariable analysis.

		AIC				Plausible	
SNP ID	**Initial Model**	**A***	**D***	**R***	**C***	**Model**	**p-value****
rs9481067	Recessive	322.4	336.5	318.4	320.2	Recessive	1.24E-06
rs10511330	Dominant	320.1	319.2	338.6	321.0	Dominant	1.40E-06
rs16822593	Dominant	320.2	319.3	338.6	321.1	Dominant	1.50E-06
rs13019215	Dominant	319.0	318.6	336.5	320.2	Dominant	8.20E-06
rs12471607	Dominant	318.9	318.6	336.4	320.1	Dominant	8.42E-06
rs716897	Dominant	323.4	321.2	337.0	323.0	Dominant	1.12E-05
rs4843335	Dominant	324.5	324.1	339.9	326.0	Dominant	1.48E-05
rs11968293	Dominant	327.2	322.4	338.9	324.4	Dominant	1.48E-05
rs4837345	Recessive	333.9	340.5	324.6	325.7	Recessive	1.97E-05
kgp10457679	Recessive	333.9	340.5	324.6	325.7	Recessive	1.97E-05
kgp4136779	Recessive	333.9	340.5	324.6	325.7	Recessive	1.97E-05
rs1075650	Recessive	334.1	340.6	324.6	325.6	Recessive	1.97E-05
rs1661281	Recessive	338.2	340.8	324.9	322.7	Recessive	2.17E-05
rs6573132[1]	Co-Dominant	334.3	328.9	339.0	325.0	Co-Dominant	2.27E-05
rs205536	Additive	321.3	330.3	326.0	323.3	Additive	2.34E-05
rs17093005[1]	Co-Dominant	334.2	328.9	338.8	325.1	Co-Dominant	2.66E-05
rs11648965	Additive	324.4	325.2	336.7	326.3	Additive	2.82E-05
rs8019850[1]	Co-Dominant	334.4	329.2	338.9	325.4	Co-Dominant	3.07E-05
rs12915222	Additive	323.5	327.9	331.9	325.5	Additive	3.19E-05
rs1189903[1]	Co-Dominant	328.0	323.4	334.2	320.9	Co-Dominant	3.47E-05
rs6596805	Recessive	331.2	339.6	325.3	327.1	Recessive	4.51E-05
rs11159673	Additive	325.5	329.4	331.1	326.8	Additive	4.53E-05
rs11047047	Recessive	328.4	338.3	325.2	327.1	Recessive	4.79E-05
rs2384298	Additive	323.3	324.0	333.6	324.5	Additive	5.14E-05
rs9809129	Dominant	319.6	321.4	335.0	321.0	Additive	6.52E-05
rs11656626[2]	Co-Dominant	331.5	337.6	326.6	328.3	Recessive	6.94E-05
rs12956191	Additive	325.5	330.0	331.2	327.4	Additive	8.18E-05
rs17712784	Dominant	327.6	326.5	340.7	328.5	Dominant	1.12E-04
rs7314811[2]	Co-Dominant	328.2	334.1	329.8	329.4	Additive	1.63E-04
rs16907305[2]	Co-Dominant	329.1	334.8	330.3	330.3	Additive	2.66E-04
rs919001	Recessive	330.4	337.4	329.7	331.0	Recessive	4.66E-04
rs11216624	Dominant	332.7	330.0	339.9	329.5	Co-Dominant	5.67E-04
rs4779810[2]	Co-Dominant	330.5	333.7	334.5	332.5	Additive	1.30E-03

*A: Additive, D: Dominant, R: Recessive, C: Co-dominant.
**p-value under the multivariable model based on the initial genetic model.
[1]: heterozygous genotype/major allele homozygous genotype.
[2]: minor allele homozygous genotype/major allele homozygous genotype.
The SNPs in bold were identified under their plausible genetic model.

Supplementary Table S7. Haploreg results for the top 10 SNPs in the common variant analysis.

SNP ID	chr	r2	D'	rs ID	GENCODE_name
rs10819474	9	0.92	0.96	rs4837345	PPP2R4
	9	0.91	0.96	rs192983	IER5L
	9	0.9	-0.97	rs944072	IER5L
	9	0.86	-0.99	rs10819473	IER5L
	9	0.94	-0.99	rs1966223	IER5L
	9	0.94	-0.99	rs1966222	IER5L
	9	0.99	1	rs12057089	IER5L
	9	1	1	rs10819474	IER5L
	9	1	1	rs10819475	IER5L
	9	1	1	rs419636	IER5L
	9	0.96	-1	rs12237274	IER5L
	9	0.9	-0.99	rs10739743	IER5L
	9	0.88	-0.99	rs4837346	IER5L
	9	0.94	-0.99	rs1556147	IER5L
	9	0.97	0.99	rs141780496	IER5L
	9	0.97	0.99	rs1075650	IER5L
	9	0.97	0.99	rs184457	IER5L
	9	0.93	0.97	rs882616	RP11-247A12.2
	9	0.93	-0.99	rs7034195	RP11-247A12.2
	9	0.91	-0.99	rs2005078	RP11-247A12.2
	9	0.83	-0.93	rs967497	RP11-247A12.2
	9	0.85	0.95	rs913264	RP11-247A12.2
	9	0.83	-0.93	rs4837347	RP11-247A12.2
	9	0.82	-0.92	rs7871824	RP11-247A12.2
rs716897	5	1	1	rs716897	RASGRF2
rs4843335	16	0.96	0.98	rs7500355	RP11-805I24.1
	16	1	1	rs4843335	RP11-805I24.1
rs4837345	9	0.8	0.94	rs9408986	PPP2R4
	9	0.8	0.94	rs71497442	PPP2R4
	9	0.8	0.94	rs4836641	PPP2R4
	9	0.8	0.94	rs1107329	PPP2R4
	9	1	1	rs4837345	PPP2R4
	9	0.97	1	rs192983	IER5L
	9	0.95	-1	rs944072	IER5L
	9	0.82	-0.97	rs10819473	IER5L
	9	0.9	-0.97	rs1966223	IER5L
	9	0.9	-0.97	rs1966222	IER5L
	9	0.91	0.95	rs12057089	IER5L
	9	0.92	0.96	rs10819474	IER5L
	9	0.92	0.96	rs10819475	IER5L

	9	0.92	0.96	rs419636	IER5L
	9	0.89	-0.97	rs12237274	IER5L
	9	0.83	-0.95	rs10739743	IER5L
	9	0.81	-0.95	rs4837346	IER5L
	9	0.87	-0.95	rs1556147	IER5L
	9	0.89	0.95	rs141780496	IER5L
	9	0.89	0.95	rs1075650	IER5L
	9	0.89	0.95	rs184457	IER5L
	9	0.86	0.93	rs882616	RP11-247A12.2
	9	0.87	-0.95	rs7034195	RP11-247A12.2
	9	0.84	-0.95	rs2005078	RP11-247A12.2
rs9481067	6	1	1	rs9481067	SLC22A16
	6	0.99	1	rs910399	SLC22A16
	6	1	1	rs761589	SLC22A16
rs13019215	2	1	1	rs13019215	CCDC141
	2	0.94	1	rs11680978	CCDC141
	2	0.93	1	rs150840830	CCDC141
	2	0.93	1	rs10930850	CCDC141
	2	0.92	0.99	rs12471607	CCDC141
rs12471607	2	0.92	0.99	rs13019215	CCDC141
	2	0.98	0.99	rs11680978	CCDC141
	2	0.98	0.99	rs150840830	CCDC141
	2	0.98	0.99	rs10930850	CCDC141
	2	1	1	rs12471607	CCDC141
rs10511330	3	0.92	0.96	rs16822588	ZBTB20
	3	0.95	0.99	rs16822593	ZBTB20
	3	0.97	0.99	rs73857113	ZBTB20
	3	1	1	rs6763403	ZBTB20
	3	1	1	rs10511330	ZBTB20
	3	0.94	0.97	rs73860251	ZBTB20
	3	0.91	0.97	rs16822606	ZBTB20
	3	0.91	0.97	rs7428451	ZBTB20
	3	0.9	0.96	rs6778079	ZBTB20
	3	0.9	0.96	rs6792964	ZBTB20
	3	0.9	0.96	rs6793257	ZBTB20
	3	0.9	0.96	rs57864250	ZBTB20
	3	0.9	0.96	rs73857603	ZBTB20
	3	0.9	0.96	rs6785090	ZBTB20
	3	0.9	0.96	rs73857605	ZBTB20
	3	0.9	0.96	rs2067756	ZBTB20
rs16822593	3	0.97	1	rs16822588	ZBTB20
	3	1	1	rs16822593	ZBTB20
	3	0.97	1	rs73857113	ZBTB20
	3	0.95	0.99	rs6763403	ZBTB20

	3	0.95	0.99	rs10511330	ZBTB20
	3	0.88	0.95	rs73860251	ZBTB20
	3	0.86	0.93	rs16822606	ZBTB20
	3	0.86	0.93	rs7428451	ZBTB20
	3	0.85	0.92	rs6778079	ZBTB20
	3	0.85	0.92	rs6792964	ZBTB20
	3	0.85	0.92	rs6793257	ZBTB20
	3	0.85	0.92	rs57864250	ZBTB20
	3	0.85	0.92	rs73857603	ZBTB20
	3	0.85	0.92	rs6785090	ZBTB20
	3	0.85	0.92	rs73857605	ZBTB20
	3	0.85	0.92	rs2067756	ZBTB20
rs11968293	6	0.9	0.98	rs10708664	SLC35F1
	6	0.88	0.99	rs6940985	SLC35F1
	6	1	1	rs11968293	SLC35F1
	6	1	1	rs1572226	SLC35F1
	6	0.85	0.97	rs72967533	16kb 3' of SLC35F1
	6	0.82	0.93	rs11153730	29kb 3' of SLC35F1

To retrieve the data, we used the default conditions and selected the CEU (Caucasian) as the population.
Chr: chromosome, r2: correlation coefficient,
D': ratio of given and min/max coefficient of linkage disequilibrium depending on allele frequencies.

Supplementary Table S8. Proteins which have reported evidence of binding to the genomic region in which kgp10457679 resides (extracted from RegulomeDB).

Location	Bound Protein
chr9:131930083-131930707	PHF8
chr9:131930124-131930734	POLR2A
chr9:131930177-131930837	SMARCB1
chr9:131930235-131930775	EP300
chr9:131930398-131930728	EP300
chr9:131930394-131930710	EP300
chr9:131930310-131930710	FOS
chr9:131930295-131930751	GTF2F1
chr9:131930295-131930751	JUND
chr9:131930321-131930661	JUN
chr9:131930369-131930709	JUN
chr9:131930291-131930687	MAX
chr9:131930308-131930772	POLR2A
chr9:131930314-131930520	POLR2A
chr9:131930373-131930729	RCOR1
chr9:131930379-131930709	RFX5
chr9:131930359-131930715	TAF1
chr9:131930275-131930711	UBTF
chr9:131930369-131930679	IRF1
chr9:131930380-131930700	JUND
chr9:131930343-131930699	MAX
chr9:131930332-131930702	MAZ
chr9:131930374-131930684	MYC
chr9:131930403-131930699	RCOR1
chr9:131930351-131930667	YY1
chr9:131930381-131930677	JUN
chr9:131930388-131930672	FOS
chr9:131930392-131930656	CEBPB
chr9:131930395-131930671	JUND
chr9:131930405-131930655	FOS
chr9:131930406-131930666	FOSL2
chr9:131930420-131930656	BACH1
chr9:131930423-131930659	ATF3
chr9:131930419-131930649	JUN
chr9:131930417-131930627	MAFF
chr9:131930376-131930666	USF2
chr9:131930451-131930567	MAFK
chr9:131930484-131930584	FOS
chr9:131930489-131930574	FOS
chr9:131930462-131930577	NFE2

chr9:131930433-131930605	FOSL2
chr9:131930433-131930610	MAFF
chr9:131930436-131930602	MAFK
chr9:131930442-131930598	MAFK
chr9:131930432-131930604	MAFK
chr9:131930210-131930654	POLR2A
chr9:131930433-131930663	FOSL1
chr9:131930443-131930622	BACH1
chr9:131930438-131930615	MAFK

Appendix C: Supplementary data for Penney *et al.*, 2018: “A Genome-wide Association Study Identifies Single Nucleotide Polymorphisms Associated with Time-to-Metastasis in Colorectal Cancer”

Supplementary Table S1. Results from the stepwise variable selection method using multivariable mixture cure model to determine the final significant baseline characteristics

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
Location (rectum vs. colon)	9.11	1.05-78.82	0.0447	0.23	0.09-0.58	0.0018
5-FU treatment status (given vs. not given/unknown)	7.15	1.37-37.26	0.0195	0.21	0.06-0.69	0.0103
Stage II (vs. Stage I)	1.25	0.12-13.47	0.8529	3.45	0.34-35.37	0.2964
Stage III (vs. Stage I)	0.45	0.02-8.75	0.5990	14.22	1.27-159.49	0.0313

OR: odds ratio for metastasis (i.e. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S2. Results from the univariable analysis conducted on the baseline characteristics using the Cox PH model to identify the factors to adjust in the multivariable analysis

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
5-FU treatment status (given vs. not given/unknown)	1.37	0.64-2.91	0.4078
Stage II (vs. Stage I)	1.92	0.71-5.17	0.1986
Stage III (vs. Stage I)	3.10	1.04-9.24	**0.0424**
Location (rectum vs. colon)	1.76	1.06-2.92	**0.0278**
BRAF V600E mutation (present vs. absent)	2.83	1.30-6.16	**0.0085**

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval. 5-FU: 5-fluorouracil

Supplementary Table S3. Demographic and clinicopathologic characteristics of the patient cohort and *larger NFCCR cohort

Variable		Number of Patients in NFCCR Cohort (n=517)[a]	% Total	Number of Patients in Sample Cohort (n=379)[a]	% Total
Sex	Female	194	37.5%	139	36.7%
	Male	323	62.5%	240	63.3%
Age	≤60	205	39.7%	157	41.4%
	60-70	212	41.0%	154	40.6%
	70<	100	19.3%	68	17.9%
Familial risk	Low	260	50.3%	196	51.7%
	Intermediate/high	241	46.6%	183	48.3%
	Unknown	15	2.9%	N/A	N/A
5-FU based treatment	5-FU treated	279	54.0%	214	56.5%
	other/no chemo	228	44.1%	159	42.0%
	Unknown	10	1.9%	6	1.6%
Stage	I	97	18.8%	81	21.4%
	II	211	40.8%	158	41.7%
	III	209	40.4%	140	36.9%
Location	Colon	321	62.1%	233	61.5%
	Rectum	196	37.9%	146	38.5%
Histology	Non-mucinous	462	89.4%	343	90.5%
	Mucinous	55	10.6%	36	9.5%
Vascular invasion	Absence	313	60.5%	242	63.9%
	Presence	167	32.3%	111	29.3%
	Unknown	37	7.2%	26	6.9%
Lymphatic invasion	Absence	306	59.2%	237	62.5%
	Presence	170	32.9%	116	30.6%
	Unknown	41	7.9%	26	6.9%
BRAF V600E mutation	Absence	433	83.8%	333	87.9%
	Presence	35	6.8%	19	5.0%
	Unknown	49	9.5%	27	7.1%

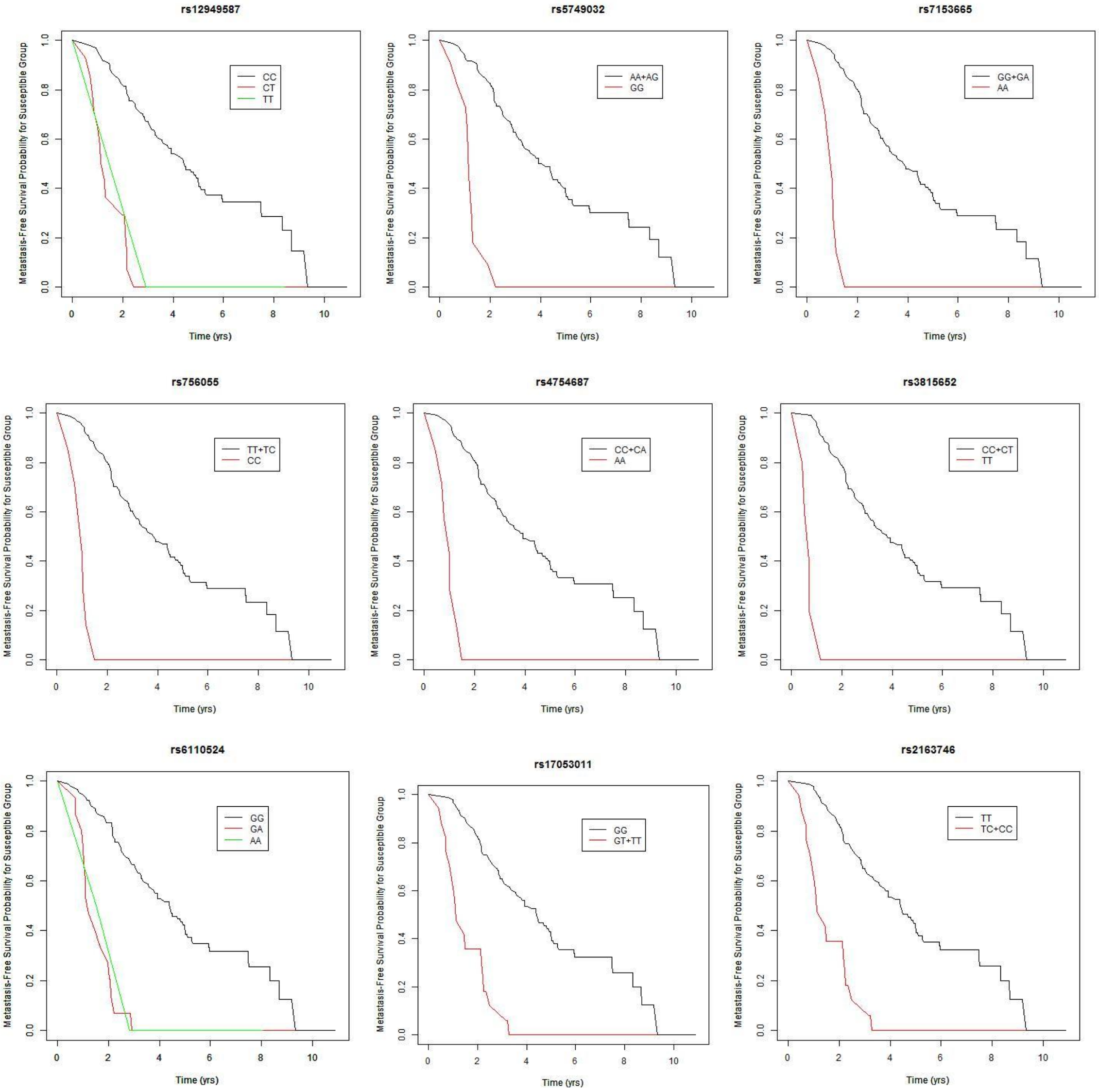
5-FU: 5-fluorouracil. *NFCCR included 750 consenting patients diagnosed with colorectal cancer between 1999 and 2003 [1,2]. From this set of patients, only the patients with MSI-L/MSS tumors and Stage I-III patients are shown in this table.

Supplementary Table S4. Genotypes significantly associated with time-to-metastasis identified in the univariable analysis using the mixture cure model

			Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Genomic Location	**Genetic Model**	**rs Number (genotypes *a* vs. *b*)**	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
22:17793969	Recessive	rs5749032 (GG vs. AA + AG)	0.73	0.35 - 1.53	0.400	9.55	4.44 - 20.55	7.70E-09
17:77361176	Co-Dominant	rs12949587 (CT vs. CC)	0.61	0.31 - 1.20	0.151	7.92	3.88 - 16.16	1.29E-08
20:15111138	Co-Dominant	rs6110524 (AG vs. GG)	0.86	0.44 - 1.70	0.665	7.56	3.75 - 15.27	1.66E-08
7:33913404	Recessive	rs3815652 (TT vs. CC + CT)	1.38	0.46 - 4.15	0.566	20.75	7.20 - 59.80	1.96E-08
14:100691178	Recessive	rs756055 (CC vs. TT + TC)	0.44	0.18-1.03	0.058	13.39	5.37 - 33.43	2.70E-08
14:100730920	Recessive	rs7153665 (AA vs. GG + AG)	0.44	0.18-1.04	0.058	13.39	5.37 - 33.44	2.70E-08
11:100430053	Recessive	rs4754687 (AA vs. CC + CA)	0.60	0.25 - 1.44	0.255	13.33	5.34 - 33.28	2.90E-08
5:155345221	Dominant	rs2163746 (CT + CC vs. TT)	0.60	0.31 - 1.15	0.124	6.45	3.29 - 12.63	5.40E-08
5:155361116	Dominant	rs17053011 (TG + TT vs. GG)	0.60	0.31 - 1.16	0.124	6.45	3.29 - 12.64	5.40E-08

OR: odds ratio for metastasis (i.e. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval

Supplementary Figure S1. Conditional survival functions for the nine SNPs identified in the univariable analysis using the mixture cure model



Under the assumptions of the mixture cure model, the population is viewed as a mixture of susceptible and non-susceptible individuals to metastasis, where susceptible refers to patients who will experience metastasis and non-susceptible individuals are long-term metastasis-free survivors who are viewed as (statistically) cured. For example, Figure 4a shows the Kaplan-Meier estimate of the survival curves of time-to-metastasis (T) for each genotype category of a specific SNP rs5749032 (i.e., Kaplan-Meier estimates of the survival function $S(t|x) = Pr(T > t|x)$ where x denotes the covariate (genotype category of the corresponding polymorphism)). On the other hand, the plots in this figure show the estimated conditional survival curves for the susceptible group under each x level (i.e., Kaplan-Meier estimates of $S_0(t|x) = Pr(T > t \mid susceptible\ group, x)$ which is the probability that the susceptible person will survive beyond a specified time t without metastasis). Hence, in Figure 4a, the probability of survival is for the population under consideration including both susceptible and non-susceptible individuals, but the plots in this figure are for the survival function of time-to-metastasis in the group of susceptible individuals. The conditional survival curves for the susceptible group are obtained from the mixture cure model $S(t|x) = Pr(T > t|x) = p(x) + (1 - p(x))S_0(t|x)$ where $p(x)$ denotes the probability of being long-term metastasis-free survivor and thus $1 - p(x)$ is the probability of being susceptible to metastasis. Hence, the curves in this figure were obtained by plugging the Kaplan-Meier estimates of $S(t|x)$ and $p(x)$ in $S_0(t|x) = (S(t|x) - p(x))/(1 - p(x))$.

Supplementary Table S5. Results from the multivariable analysis of rs12949587 using the mixture cure model under the co-dominant genetic model

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs12949587 (CT vs. CC)	0.66	0.32-1.37	0.2614	7.56	3.44-16.61	4.63E-07
rs12949587 (TT vs. CC)	0.48	0.04-5.27	0.5486	2.21	0.18-27.82	0.5397
Location (rectum vs. colon)	2.11	1.08-4.11	0.0292	0.47	0.23-0.95	0.0351
5-FU treatment (given vs. not given/unknown)	1.96	0.76-5.02	0.1617	0.62	0.20-1.97	0.4197
Stage II (vs. Stage I)	2.51	0.82-7.73	0.1076	0.81	0.21-3.16	0.7659
Stage III (vs. Stage I)	2.99	0.80-11.20	0.1051	1.96	0.37-10.26	0.4269

OR: odds ratio for metastasis (i.e. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S6. Results from the multivariable analysis of rs6110524 using the mixture cure model under the co-dominant genetic model

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs6110524 (AG vs. GG)	0.95	0.44-2.04	0.8870	4.80	2.00-11.53	0.0005
rs6110524 (AA vs. GG)	1.55	0.18-13.39	0.6886	5.24	0.93-29.65	0.0608
Location (rectum vs. colon)	2.33	1.05-5.15	0.0366	0.40	0.17-0.95	0.0379
5-FU treatment (given vs. not given/unknown)	2.24	0.78-6.48	0.1352	0.40	0.12-1.39	0.1491
Stage II (vs. Stage I)	2.24	0.70-7.17	0.1728	1.20	0.31-4.68	0.7958
Stage III (vs. Stage I)	2.47	0.57-10.74	0.2274	3.18	0.52-19.36	0.2094

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S7. Results from the multivariable analysis of rs17053011 using the mixture cure model under the dominant genetic model

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs17053011 (TG + TT vs. GG)	0.49	0.23-1.07	0.0746	9.65	3.67-25.37	4.29E-06
Location (rectum vs. colon)	2.33	1.12-4.85	0.0230	0.44	0.21-0.92	0.0296
5-FU treatment (given vs. not given/unknown)	2.24	0.73-6.91	0.1606	0.68	0.17-2.79	0.5935
Stage II (vs. Stage I)	2.16	0.55-8.46	0.2706	1.14	0.21-6.09	0.8791
Stage III (vs. Stage I)	1.86	0.36-9.58	0.4568	4.78	0.69-33.40	0.1143

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S8. Results from the multivariable analysis of rs2163746 using the mixture cure model under the dominant genetic model

	Logistic regression model for metastasis probability			*Proportional hazards model for time-to-metastasis*		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs2163746 (CT + CC vs. TT)	0.49	0.23-1.07	0.0746	9.65	3.67-25.37	4.29E-06
Location (rectum vs. colon)	2.33	1.12-4.85	0.0230	0.44	0.21-0.92	0.0296
5-FU treatment (given vs. not given/unknown)	2.24	0.73-6.91	0.1606	0.68	0.17-2.79	0.5935
Stage II (vs. Stage I)	2.16	0.55-8.46	0.2706	1.14	0.21-6.09	0.8791
Stage III (vs. Stage I)	1.86	0.36-9.58	0.4568	4.78	0.69-33.40	0.1143

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S9. Results from the multivariable analysis of rs3815652 using the mixture cure model under the recessive genetic model

	Logistic regression model for metastasis probability			*Proportional hazards model for time-to-metastasis*		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs3815652 (TT vs. CC + CT)	0.59	0.13-2.65	0.4878	12.97	3.26-51.66	0.0003
Location (rectum vs. colon)	4.17	1.28-13.62	0.0179	0.28	0.13-0.59	0.0008
5-FU treatment (given vs. not given/unknown)	4.88	0.82-28.97	0.0810	0.20	0.05-0.77	0.0197
Stage II (vs. Stage I)	2.27	0.62-8.34	0.2171	1.80	0.43-7.60	0.4218
Stage III (vs. Stage I)	1.13	0.14-9.22	0.9070	7.67	1.10-53.51	0.0398

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S10. Results from the multivariable analysis of rs4754687 using the mixture cure model under the recessive genetic model

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs4754687 (AA vs. CC + CA)	0.51	0.18-1.43	0.2012	8.13	2.59-25.53	0.0003
Location (rectum vs. colon)	2.61	1.06-6.39	0.0366	0.34	0.15-0.75	0.0082
5-FU treatment (given vs. not given/unknown)	2.35	0.70-7.85	0.1656	0.48	0.11-2.10	0.3314
Stage II (vs. Stage I)	2.28	0.68-7.64	0.1799	1.20	0.28-5.18	0.8110
Stage III (vs. Stage I)	2.38	0.48-11.92	0.2901	3.09	0.40-23.70	0.2783

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S11. Results from the multivariable analysis of rs5749032 using the mixture cure model under the recessive genetic model

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs5749032 (GG vs. AA + AG)	0.38	0.14-1.07	0.0661	15.86	6.83-36.83	1.28E-10
Location (rectum vs. colon)	4.01	1.65-9.71	0.0021	0.28	0.15-0.52	4.56E-05
5-FU treatment (given vs. not given/unknown)	5.20	1.30-20.71	0.0194	0.14	0.05-0.41	0.0003
Stage II (vs. Stage I)	2.57	0.78-8.51	0.1214	1.60	0.46-5.63	0.4624
Stage III (vs. Stage I)	1.18	0.22-6.38	0.8439	8.50	1.82-39.64	0.0064

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S12. Results from the multivariable analysis of rs756055 using the mixture cure model under the recessive genetic model

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs756055 (CC vs. TT + TC)	0.28	0.10-0.82	0.0204	7.58	2.53-22.65	0.0003
Location (rectum vs. colon)	2.47	1.02-5.98	0.0442	0.36	0.17-0.78	0.0097
5-FU treatment (given vs. not given/unknown)	2.72	0.77-9.64	0.1208	0.41	0.10-1.79	0.2389
Stage II (vs. Stage I)	2.17	0.63-7.48	0.2177	1.38	0.29-6.59	0.6845
Stage III (vs. Stage I)	2.13	0.40-11.35	0.3771	3.85	0.45-33.07	0.2185

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S13. Results from the multivariable analysis of rs7153665 using the mixture cure model under the recessive genetic model

	Logistic regression model for metastasis probability			***Proportional hazards model for time-to-metastasis***		
Variable (*a* vs. *b*)	**OR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
rs7153665 (AA vs. GG + AG)	0.28	0.10-0.82	0.0204	7.58	2.53-22.65	0.0003
Location (rectum vs. colon)	2.47	1.02-5.98	0.0442	0.36	0.17-0.78	0.0097
5-FU treatment (given vs. not given/unknown)	2.72	0.77-9.64	0.1208	0.41	0.10-1.79	0.2389
Stage II (vs. Stage I)	2.17	0.63-7.48	0.2177	1.38	0.29-6.59	0.6845
Stage III (vs. Stage I)	2.13	0.40-11.35	0.3771	3.85	0.45-33.07	0.2185

OR: odds ratio for metastasis (ie. probability of being in the susceptible group). OR compares metastasis proportion in subgroup *a* with that in subgroup *b*. HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S14. Results for all significant SNPs in the univariable Cox PH analysis and subsequent multivariable results

		Univariable			Multivariable		
Genomic Location	**rs Number (Genotype)**	**HR**	**95% CI**	**p-value**	**HR**	**95% CI**	**p-value**
20:16189263	rs2327990 (TT)	21.97	8.42 - 57.33	2.74E-10	22.58	8.32-61.31	9.59E-10
3:134513356	rs11918092 (CC)	216.98	35.64 - 1321.13	5.32E-09	535.33	63.20-4534.30	8.23E-09
3:134515336	rs3732568 (AA)	216.98	35.64 - 1321.13	5.32E-09	535.33	63.20-4534.30	8.23E-09
3:59930672	rs2366964 (CC)	41.19	11.81 - 143.66	5.40E-09	56.53	14.98-213.26	2.59E-09
2:175205513	rs7582977 (CC)	134.32	25.76 - 700.33	6.02E-09	82.61	14.50-470.67	6.63E-07
13:48118782	rs9534678 (AA)	133.60	25.62 - 696.59	6.26E-09	83.96	14.71-479.13	6.17E-07
2:86015121	rs13402783 (GG)	20.91	7.47 - 58.50	6.94E-09	13.03	4.50-37.78	2.25E-06
2:86013029	rs13386681 (TT)	20.79	7.47 - 58.50	7.40E-09	12.86	4.43-37.27	2.57E-06
2:6769988	rs1563948 (AA)	34.43	10.35 - 114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
2:6773920	rs11692570 (TT)	34.43	10.35 - 114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
2:6777992	rs2219613 (TT)	34.43	10.35 - 114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
2:6779277	rs11694697 (TT)	34.43	10.35 - 114.58	7.97E-09	33.97	9.57-120.54	4.87E-08
5:148172928	rs9285673 (CC)	36.70	10.53 - 127.95	1.56E-08	19.47	5.41-70.13	5.60E-06
15:89420974	rs17201864 (TT)	19.06	6.86 - 52.98	1.60E-08	11.01	3.76-32.24	1.20E-05
9:119519588	rs1372330 (AA)	36.51	10.47 - 127.34	1.67E-08	27.15	7.66-96.27	3.20E-07
6:91187510	rs1145724 (GG)	30.76	9.27 - 102.03	2.14E-08	36.43	10.21-129.93	3.00E-08
4:53893156	rs17082301 (AA)	129.98	23.29 - 725.52	2.89E-08	81.63	8.85-753.27	0.0001
1:190131750	rs10920654 (TT)	76.85	16.51 - 357.84	3.16E-08	32.48	5.86-180.04	6.78E-05
10:98422896	rs1023741 (CC)	18.64	6.57 - 52.84	3.76E-08	17.77	5.07-62.25	6.87E-06
4:14296300	rs1426107 (AA)	28.08	8.53 - 92.45	4.13E-08	14.10	2.70-73.53	0.0017
18:40691675	rs3861289 (AA)	19.16	6.62 - 55.40	5.04E-08	8.73	2.58-29.54	0.0005
17:7396267	rs4265880 (AA)	98.28	18.72 - 515.88	5.86E-08	64.74	11.27-371.86	2.93E-06
17:7397043	rs4239258 (TT)	98.28	18.72 - 515.88	5.86E-08	64.74	11.27-371.86	2.93E-06
17:7404991	rs2228130 (TT)	98.28	18.72 - 515.88	5.86E-08	64.74	11.27-371.86	2.93E-06
17:7418109	rs9989479 (AA)	98.28	18.72 - 515.88	5.86E-08	64.74	11.27-371.86	2.93E-06

HR: hazard ratio for time to metastasis among susceptible patients. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S15. Results from the multivariable analysis of rs2327990 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs2327990 (TT vs. CC + CT)	22.58	8.32-61.31	9.59E-10
Location (rectum vs. colon)	1.69	1.01-2.80	0.044
5-FU treatment (given vs. not given/unknown)	1.32	0.62-2.83	0.469
Stage II (vs. Stage I)	2.14	0.78-5.87	0.139
Stage III (vs. Stage I)	3.46	1.13-10.54	0.029
BRAF V600E mutation (present vs. absent)	3.01	1.38-6.56	0.005

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S16. Results from the multivariable analysis of rs3732568 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs3732568 (AA vs. CC + CA)	535.33	63.20-4534.30	8.23E-09
Location (rectum vs. colon)	1.76	1.06-2.93	0.029
5-FU treatment (given vs. not given/unknown)	1.55	0.71-3.38	0.269
Stage II (vs. Stage I)	1.69	0.62-4.63	0.306
Stage III (vs. Stage I)	2.75	0.91-8.31	0.074
BRAF V600E mutation (present vs. absent)	2.94	1.35-6.41	0.007

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S17. Results from the multivariable analysis of rs2219613 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs2219613 (TT vs. CC + CT)	33.97	9.57-120.54	4.87E-08
Location (rectum vs. colon)	1.96	1.17-3.30	0.011
5-FU treatment (given vs. not given/unknown)	1.29	0.60-2.75	0.513
Stage II (vs. Stage I)	2.00	0.74-5.40	0.173
Stage III (vs. Stage I)	3.07	1.02-9.21	0.046
BRAF V600E mutation (present vs. absent)	3.16	1.44-6.91	0.004

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S18. Results from the multivariable analysis of rs11692570 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs11692570 (TT vs. CC + CT)	33.97	9.57-120.54	4.87E-08
Location (rectum vs. colon)	1.96	1.17-3.30	0.011
5-FU treatment (given vs. not given/unknown)	1.29	0.60-2.75	0.513
Stage II (vs. Stage I)	2.00	0.74-5.40	0.173
Stage III (vs. Stage I)	3.07	1.02-9.21	0.046
BRAF V600E mutation (present vs. absent)	3.16	1.44-6.91	0.004

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S19. Results from the multivariable analysis of rs11918092 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs11918092 (CC vs. AA + AC)	535.33	63.20-4534.30	8.23E-09
Location (rectum vs. colon)	1.76	1.06-2.93	0.029
5-FU treatment (given vs. not given/unknown)	1.55	0.71-3.38	0.269
Stage II (vs. Stage I)	1.69	0.62-4.63	0.306
Stage III (vs. Stage I)	2.75	0.91-8.31	0.074
BRAF V600E mutation (present vs. absent)	2.94	1.35-6.41	0.007

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S20. Results from the multivariable analysis of rs1145724 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs1145724 (GG vs AA + AG)	36.43	10.21-129.93	3.00E-08
Location (rectum vs. colon)	1.85	1.12-3.06	0.016
5-FU treatment (given vs. not given/unknown)	1.23	0.58-2.62	0.588
Stage II (vs. Stage I)	1.91	0.71-5.17	0.203
Stage III (vs. Stage I)	3.33	1.12-9.92	0.031
BRAF V600E mutation (present vs. absent)	3.07	1.41-6.70	0.005

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S21. Results from the multivariable analysis of rs11694697 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs11694697 (TT vs CC + CT)	33.97	9.57-120.54	4.87E-08
Location (rectum vs. colon)	1.96	1.17-3.30	0.011
5-FU treatment (given vs. not given/unknown)	1.29	0.60-2.75	0.513
Stage II (vs. Stage I)	2.00	0.74-5.40	0.173
Stage III (vs. Stage I)	3.07	1.02-9.21	0.046
BRAF V600E mutation (present vs. absent)	3.16	1.44-6.91	0.004

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S22. Results from the multivariable analysis of rs1563948 using the Cox PH model under the recessive genetic model

Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs1563948 (AA vs GG + GA)	33.97	9.57-120.54	4.87E-08
Location (rectum vs. colon)	1.96	1.17-3.30	0.011
5-FU treatment (given vs. not given/unknown)	1.29	0.60-2.75	0.513
Stage II (vs. Stage I)	2.00	0.74-5.40	0.173
Stage III (vs. Stage I)	3.07	1.02-9.21	0.046
BRAF V600E mutation (present vs. absent)	3.16	1.44-6.91	0.004

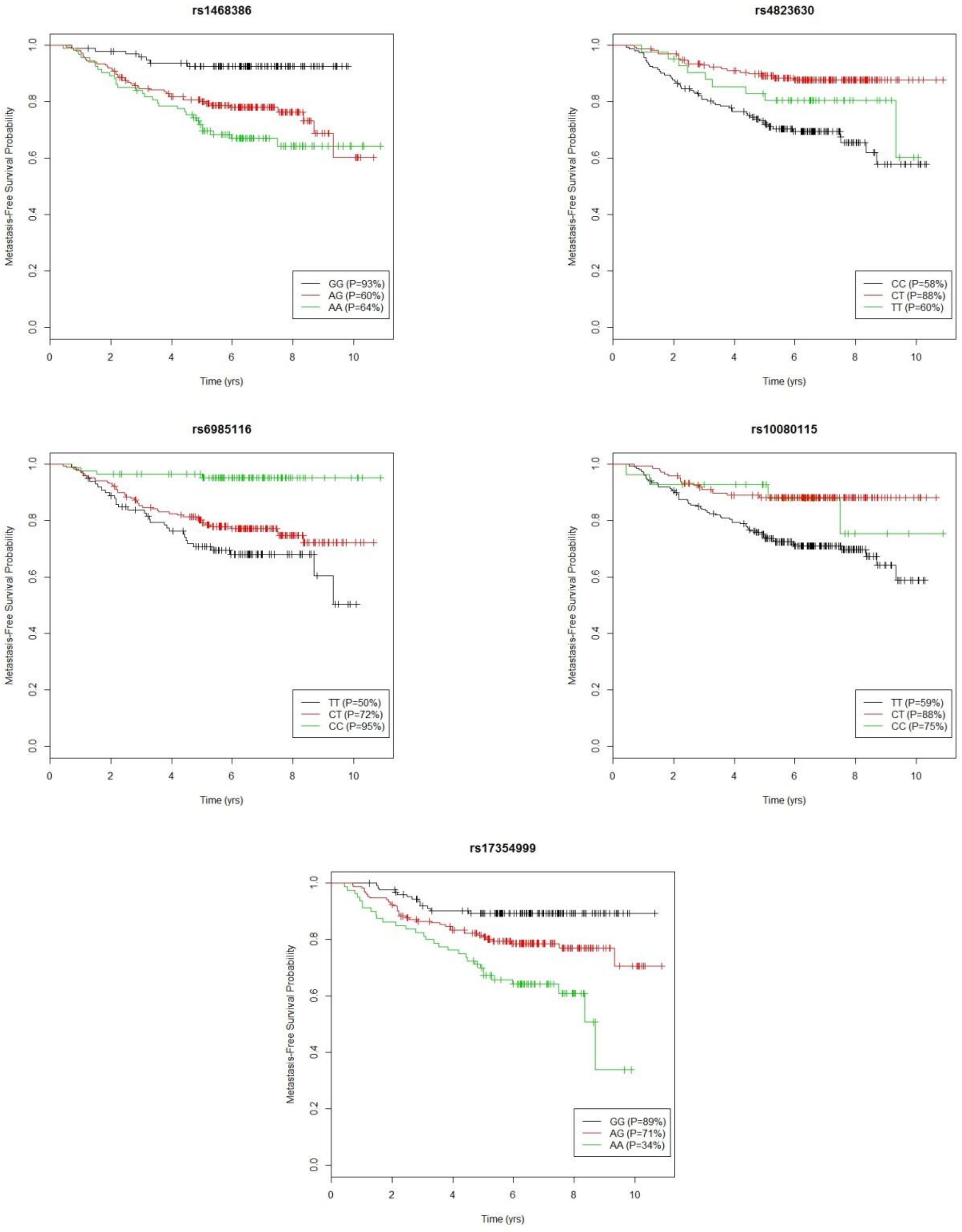
HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Table S23. Results from the multivariable analysis of rs2366964 using the Cox PH model under the recessive genetic model

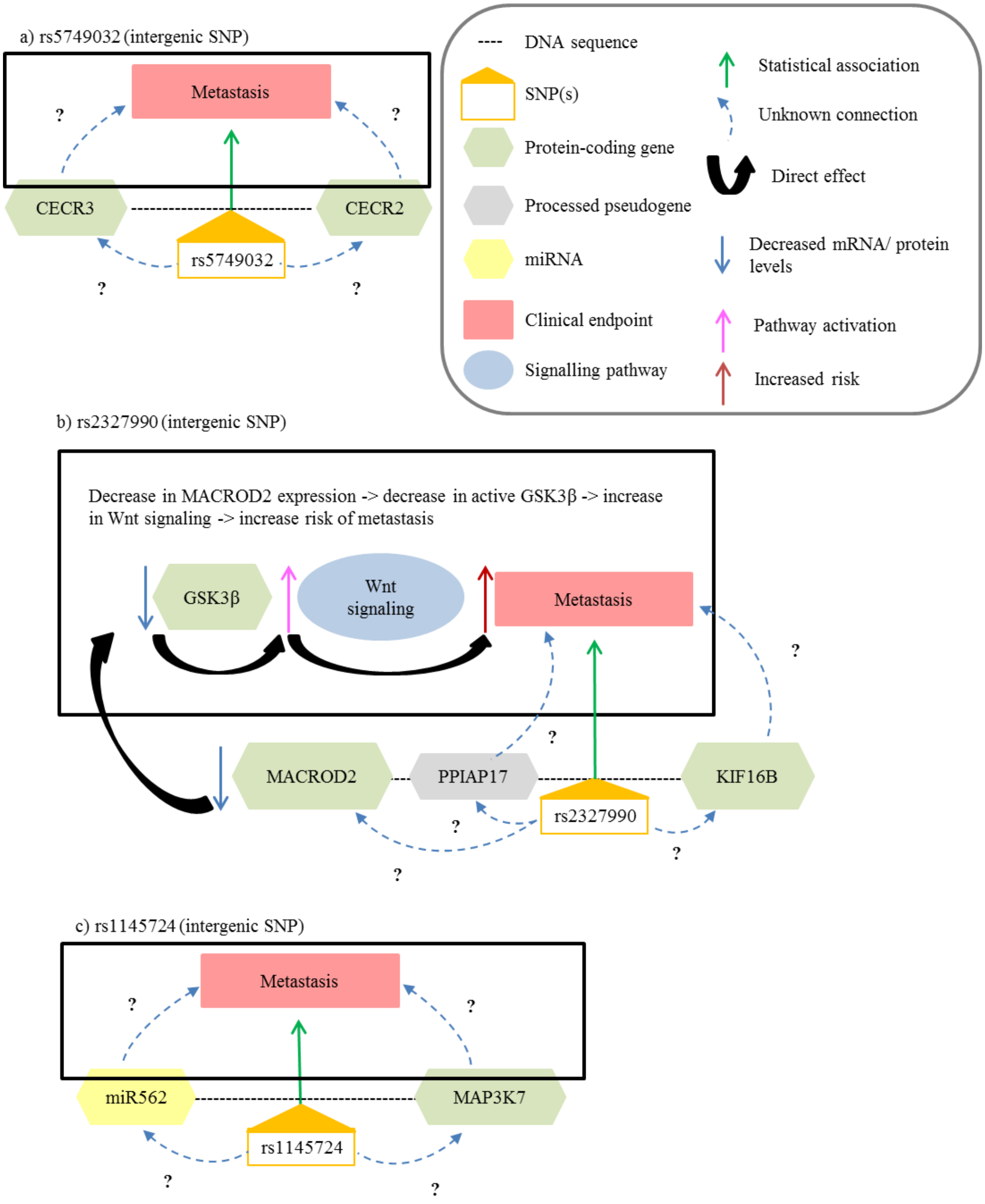
Variable (*a* vs. *b*)	**HR**	**95% CI**	**p-value**
rs2366964 (CC vs TT + TC)	56.53	14.98-213.26	2.59E-09
Location (rectum vs. colon)	1.72	1.03-2.86	0.038
5-FU treatment (given vs. not given/unknown)	1.45	0.68-3.12	0.336
Stage II (vs. Stage I)	1.69	0.62-4.59	0.307
Stage III (vs. Stage I)	2.91	0.97-8.72	0.057
BRAF V600E mutation (present vs. absent)	2.96	1.36-6.45	0.006

HR: hazard ratio for time to metastasis among susceptible patients. HR compares metastasis rate in subgroup *a* with that in subgroup *b* among those who are susceptible to metastasis. CI: confidence interval; 5-FU: 5-fluorouracil

Supplementary Figure S2. Survival curves for SNPs with the strongest association to risk of metastasis in the mixture cure model



Supplementary Figure S3. Known and hypothesized links between the intergenic SNPs, nearby genes, and the risk of metastasis



All SNPs except rs1145724, are discussed in Discussion section of the manuscript. The intergenic SNP rs1145724 was identified by the Cox PH model as significantly associated with time to metastasis. According to UCSC genome browser [3] this SNP is flanked by a miRNA, *miR562*, and a mitogen-activated protein kinase gene, *MAP3K7*. There is no scientific literature linking *miR562* to colorectal cancer. MAP3K7, on the other hand, has been shown to be linked to colorectal cancer in several studies [4-6]. MAP3K7 (TAK1) mediates signal transduction in several pathways, including negative regulation of Wnt signaling [7]. However, at the present time there is no known connection between this SNP, these genes, or colorectal cancer metastasis. It is also possible that the SNPs identified in this study may have long-distance regulatory functions.

Appendix C References

1. Wish TA, Hyde AJ, Parfrey PS, et al. Increased cancer predisposition in family members of colorectal cancer patients harboring the p.V600E BRAF mutation: A population-based study. *Cancer Epidemiol Biomarkers Prevent*. 2010;19(7):1831-1839.

2. Woods MO, Younghusband HB, Parfrey PS, et al. The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut*. 2010;59(10):1369-1377.

3. Kent WJ, Sugnet C,W., Furey TS, et al. The human genome browser at UCSC. *Genome Res.* 2002;12(6):996-1006.

4. Slattery ML, Lundgreen A, Wolff RK. Dietary influence on MAPK-signaling pathways and risk of colon and rectal cancer. *Nutr Cancer*. 2013;65(5):729-738.

5. Takahashi H, Jin C, Rajabi H, et al. Muc1-c activates the tak1 inflammatory pathway in colon cancer. *Oncogene*. 2015;34(40):5187-5197.

6. Singh A, Sweeney MF, Yu M, et al. TAK1 (MAP3K7) inhibition promotes apoptosis in KRAS-dependent colon cancers. *Cell*. 2012;148(4):639-650.

7. Behrens J. Cross-regulation of the wnt signalling pathway: A role of MAP kinases. *J Cell Sci*. 2000;113(6):911.