Genetic and clinico-demographic factors with and without time-varying associations with survival outcomes in colorectal cancer

By:

© Yajun Yu

A Thesis submitted to the School of Graduate Studies

in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

Biomedical Sciences (Human Genetics), Faculty of Medicine,

Memorial University of Newfoundland

May 2022

ABSTRACT

Predicting the risk of outcomes (e.g., death, tumor recurrence/metastasis) as well as when a patient has such high risks are important, as this information can guide the disease management and patient care in colorectal cancer. However, current established prognostic markers in colorectal cancer predicting such risks are not sufficient to stratify patients into appropriate risk groups, suggesting the need for additional prognostic markers. In addition, there are no established prognostic markers that can tell whether a patient has high outcome risk in different time frames following diagnosis. The main aim of this research was to examine the relationship between genetic/clinico-demographic factors and clinical outcomes in colorectal cancer and to identify factors that associated with patient outcomes with or without time-varying associations.

To this end, I examined ~4.7 million SNPs and hundreds of CNVs/INDELs for their associations with outcomes in colorectal cancer. My results showed that a number of CNVs/INDELs in *BRM*-741, *TGFBR3*, *FILIP1L*, *STEAP2*, *RP11-143P4.2*, and a SNP (rs7314075) in *WBP11* are associated with time to outcome events in colorectal cancer. Further, two CNVs in *PDLIM3* and *GUSBP1* were identified to be associated with patient outcomes within the first ~3 years post-diagnosis, but not after that (i.e., candidate early-outcome markers); another three variants (rs817090, rs11064732, and rs200143895) were found to be associated with patient outcomes after 5 years post-diagnosis, but not before that (i.e., candidate late-outcome markers). I also examined a set of baseline clinico-demographic factors, where a number of them were identified to be associated with outcomes in colorectal cancer, including those that have time-varying associations. By investigating a long follow up data, I was also able to examine the long term survival characteristics in this disease.

This comprehensive research described a detailed picture of genetic and clinico-demographic factors with and without time-varying associations with clinical outcomes in colorectal cancer. It provided a set of variants/factors as candidate markers predicting outcome risks for colorectal cancer patients, including candidate early- or late-outcome markers. These factors, once their prognostic value is validated, can be used to guide patients' treatment and clinical care, and improve their survival times.

GENERAL SUMMARY

Each year, colorectal cancer causes more than 9,000 deaths in Canada. Many colorectal cancer patients also experience recurrence after their diagnosis or treatment. To prevent the recurrence of tumors and to improve patients' survival, using specific types of indicators in clinical decision-making can help. These types of indicators can tell who is likely to have recurrence or die early. For those patients who have such risks, doctors can provide personalized treatment plans. This way, patients can have better outcomes. However, currently, there are only a small number of such indicators used in clinic for colorectal cancer patients.

In this thesis research, I aimed to identify such indicators. I examined ~4.7 million genetic factors. Genetic factors are DNA units or segments that differ between individuals. I found a set of genetic factors that can be such indicators (for example, genetic factors in the *BRM*, *TGFBR3*, and *WBP11* genes). I also examined a number of clinical factors. I found some clinical factors that can also be such indicators (for example, tumor location). Further, some factors that I found can also tell when a patient is likely to have recurrence or die. For example, some factors can tell whether a patient is likely to have recurrence after 5 years following diagnosis. This information is important, as it can help improve doctors' treatment plan for patients.

Overall, my thesis research found a number of genetic and clinical factors that are possible indicators to tell whether a patient is likely to have recurrence or die early. Some factors can also tell when a patient is likely to have such events. If such findings are replicated in other studies, these factors can be used to help to improve patients' survival.

CO-AUTHORSHIP STATEMENT

In this thesis research, I (the thesis author: Yajun Yu) helped with study designs and methodologies, performed statistical analyses, interpreted the results, drafted the original manuscripts, and reviewed and revised the manuscripts for studies described in Chapters 2, 4, and 5; I also performed the survival analysis examining the associations between CNVs/INDELs with survival outcomes in colorectal cancer, interpreted the results of survival analysis, and helped draft the manuscript for the Chapter 3 study.

Others also contributed to this thesis research. Dr. Sevtap Savas (my supervisor) contributed to all four studies in this thesis research. She conceived and led the studies, reviewed and revised the manuscripts for all studies, helped collect the patient-related data (Chapter 4 study), and helped draft the manuscript (Chapter 3 study). Dr. Yildiz E. Yilmaz helped with statistical methodologies, and she conceptualized and led the Chapter 5 study with Dr. Sevtap Savas. Staff members in the Newfoundland Familial Colorectal Cancer Registry (NFCCR) - Drs. Patrick Parfrey, Jane Green, and Elizabeth Dicks contributed to my thesis research by collecting patient-related data that were used in my studies. Dr. William Pollett also helped with the collection of patient-related data (Chapter 4 study). Salem Werdyani provided the CNV/INDEL data, which were examined in studies described in Chapters 3 and 5. Megan Carey helped with updating the follow-up data of the NFCCR cohort (Chapter 4). Dr. Dangxiao Cheng performed the DNA genotyping and drafted the methods part of the manuscript, reviewed and revised the manuscript (Chapter 2 study). Drs. Jingxiong Xu, Konstantin Shestopaloff, and Wei Xu contributed to the Chapter 3 study; they performed the initial quality control and population stratification analyses on the patient cohort.

ACKNOWLEDGEMENTS

There are so many people I would like to thank at this particular moment of my life. First of all, many thanks go to my supervisor, Dr. Sevtap Savas, for offering me this great opportunity to study in the Human Genetics program, and for supporting me not only in research, but also in so many other ways. As a supervisor, you are professional, enthusiastic, and always supportive. I really enjoyed my time as a trainee in your lab, and this definitely will become one of my precious memories. I am so grateful for everything that you have done for me.

Many thanks also go to my supervisory committee members, Dr. Yildiz E. Yilmaz and Dr. Guangju Zhai. Thank you for your kind guidance, suggestions, and support during my whole journey of PhD study, and thank you for writing reference letters to support my scholarship applications.

Thanks to the lovely lab colleagues, Georgia Skardasi, Salem Werdyani, Aaron Curtis, and Megan Carey, for those great times with you. A special thank goes to Georgia, you have helped me so much for setting up my study and life in St. John's; Michelle Penney, you are always helpful and such a good friend.

Alison Southerland, my best friend in Genetics, thanks a bunch for everything. You are always energetic and optimistic, and of course, very helpful. I really appreciate all those coffee chats, the suggestions and help you gave to me when I tried to understand the culture and start a life in Canada, the times that we play softball and football, and so many other things. Tammy Benteau, you are always encouraging, thank you for your kind words when I was down. Daniel Evans and Justin Pater, thank you for your suggestions for my future career.

Of course, Deborah Quinlan, the secretary in the Genetics Program, thank you for your contributions and support to the department and everyone in Genetics, including me. Thank you and others for taking care of my precious plants during the COVID-19 pandemic. Many thanks also go to other people in Genetics. I always feel so great to be with you all. A huge thanks to staff in the Research & Graduate Studies (RGS), Ann Dorward, Jules Dore, Amy Carroll, Rhonda Roebotham, Paula Browne, Janelle Skeard, Angela Dunn, and others, for your support in relation to research programs, scholarship opportunities, and finance management. Special thanks to Amy Carroll and others for helping me with the extensions of my study permits and visa. I would also like to thank many members, especially Dwayne Hart and Mitch Sturge, in the Centre for Health Informatics and Analytics (CHIA) for their support in installing software in CHIA that needed for my Chapter 5 study and for their work in maintaining the CHIA platform.

Special thanks also go to organizations that have financially supported me in the past several years of my PhD study, including the Beatrice Hunter Cancer Research Institute/Terry Fox Research Institute (BHCRI/TFRI), the Translational and Personalized Medicine Initiative (TPMI)/NL SUPPORT, and Memorial University (School of Graduate Studies and Faculty of Medicine). Your support allowed me to focus on my research study without any financial burden. In addition, funders and committees of Dr. Roger C. Green Graduate Scholarship, A.G. Hatcher Memorial Scholarship, Dr. Angus J. Neary Genetics Scholarship, Human Genetics Graduate Student Award, and Dean's Building a Healthy Tomorrow Award, many thanks to you for your generosity, kindness, and trust. Other than these, I would also like to thank faculty members in Human Genetics and other programs in or out of Memorial University for awarding me the best presentation awards (in 2017 and 2019) in the Human Genetics program.

Most importantly, my beloved parents, grandmas and grandpas (who passed away, but you are always in my heart), and many other members of the family, thank you so much for your selfless love, deep understanding, and great support throughout the past 6 years, without you I can never be what I am today. You are the source of my strength and the harbor of my heart. I love you all, as always, as forever.

To many others who were not mentioned above but gave me help and support, directly or indirectly, thank you all and much appreciated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

3D	3 dimensional
5-FU	5-fluorouracil
AFT	Accelerated failure time
AIC	Akaike Information Criterion
AJCC/UICC	American Joint Committee on Cancer/International Union Against Cancer
ASCO	American Society of Clinical Oncology
ASCRS	American Society of Colon and Rectal Cancer Surgeons
BMI	Body mass index
BRM	Brahma
BRR	Bannayan-Ruvalcaba-Riley
CEA	Carcinoembryonic antigen
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection
CEUS	Contrast-enhanced ultrasonography
CHB	Han Chinese in Beijing
CI	Confidence interval
CIMP	CpG island methylator phenotype

CIMP-H	CpG island methylator phenotype-high
CIMP-L	CpG island methylator phenotype-low
CIN	Chromosomal instability
CMS	Consensus molecular subtypes
CN	Copy number
CNV	Copy number variation
Del	Deletion
DFS	Disease-free survival
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
DSS	Disease-specific survival
EFS	Event-free survival
EGFR	Epidermal growth factor receptor
eQTL	Expression quantitative trait locus
ERUS	Endorectal ultrasonography
ESMO	European Society for Medical Oncology
EUR	European
FAP	Familial adenomatous polyposis
FCCX	Familial colorectal cancer type X

FDA	Food and Drug Administration
FDR	First-degree relative
FIT	Fecal immunochemical test
GDC	Genomic Data Commons
gFOBT	Guaiac fecal occult blood test
GWAS	Genome-wide association study
HMPS	Hereditary mixed polyposis syndrome
HNPCC	Hereditary nonpolyposis colorectal cancer
HR	Hazard ratio
HREA	Health Research Ethics Authority
HREB	Health Research Ethics Board
HWE	Hardy-Weinberg equilibrium
IBD	Inflammatory bowel disease
INDEL (or indel)	Insertion/deletion
Ins	Insertion
JPT	Japanese in Tokyo
KM	Kaplan-Meier
LD	Linkage disequilibrium
LDL-C	Low-density lipoprotein cholesterol

LOH	Loss of heterozygosity
MAF	Minor allele frequency
MAP	MUTYH-associated polyposis
MEF-2	Myocyte enhancer factor-2
MFS	Metastasis-free survival
MMR	Mismatch repair
MSI	Microsatellite instability
MSI-H	Microsatellite instability high
MSI-L	Microsatellite instability low
MSS	Microsatellite stable
NA	Not applicable
NAHR	Non-allelic homologous recombination
NCCN	National Comprehensive Cancer Network
NFCCR	Newfoundland Familial Colorectal Cancer Registry
NL	Newfoundland and Labrador
NLCHI	Newfoundland and Labrador Centre for Health Informatics
OR	Odds ratio
OS	Overall survival
PFS	Progression-free survival

PH	Proportional hazards
PJS	Peutz-Jeghers syndrome
PVI	Peritumoral vascular invasion
QC	Quality control
RFS	Recurrence-free survival or Relapse-free survival
RMFS	Recurrence/metastasis-free survival
SBR	Scarff-Bloom-Richardson
SD	Segmental duplication
SNP	Single nucleotide polymorphism
SWI/SNF	Switch/Sucrose non-fermentable
TAD	Topologically associating domain
TCGA	The Cancer Genome Atlas
TF	Transcription factor
UCSC	University of California, Santa Cruz
UTF	Uracil-tegafur
UTR	Untranslated region
YRI	Yoruba in Ibadan

RESEARCH OUTPUTS AND AWARDS/ RECOGNITION DURING THE PHD PROGRAM

Publications

1. **Yajun Yu**, Salem Werdyani, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. A comprehensive analysis of SNPs and CNVs identifies novel markers associated with disease outcomes in colorectal cancer. *Molecular Oncology* 2021, doi: 10.1002/1878-0261.13067.

2. Aaron Curtis, **Yajun Yu**, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. SNP-SNP interactions and risk of clinical outcomes in colorectal cancer: comparison of two MDR-based software and applications to SNPs from the MMP and VEGF family member genes (submitted to *Scientific Reports*).

3. **Yajun Yu**, Megan Carey, William Pollett, Jane Green, Elizabeth Dicks, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. *BMC Medicine* 2019, 17(1): 150.

4. **Yajun Yu**, Dangxiao Cheng, Patrick Parfrey, Geoffrey Liu, Sevtap Savas. Two functional indel polymorphisms in the promoter region of the Brahma gene (*BRM*) and disease risk and progression-free survival in colorectal cancer. *PLoS One*, 2018, 13(6): e0198873.

5. Salem Werdyani, **Yajun Yu**, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Patrick Parfrey, Yildiz Yilmaz, Sevtap Savas. Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. *Cancer Medicine*, 2017, 6(6): 1220–1232.

Abstracts

1. **Yajun Yu**, Salem Werdyani, Megan Carey, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. A comprehensive analysis of SNPs and CNVs in relation to survival outcomes in colorectal cancer. *The 2021 Canadian Cancer Research Conference (CCRC)*, November 08-11, 2021, online conference (***oral presentation***).

2. Aaron Curtis, **Yajun Yu**, Megan Carey, Patrick Parfrey, Yildiz Yilmaz, Sevtap Savas. Application of GMDR and Cox-MDR data-reduction methods to identify potential multi-SNP interactions associated with survival times in colorectal cancer patients. *The 2021 Canadian Cancer Research Conference (CCRC)*, November 08-11, 2021, online conference (***oral presentation*** *by Aaron Curtis*).

3. Shloka Negi, **Yajun Yu**, Sevtap Savas. Gene expression profiles, biological pathways, and disease characteristics associated with BRM expression levels in lung adenocarcinoma tumors. *The 2021 Canadian Cancer Research Conference (CCRC)*, November 08-11, 2021, online conference (***poster presentation*** *by Shloka Negi*).

4. Aaron Curtis, **Yajun Yu**, Megan Carey, Patrick Parfrey, Yildiz Yilmaz, Sevtap Savas. Application of data reduction methods to identify potential interactions between genetic markers in colorectal cancer outcomes. 2021 *Aldrich Conference*, August 16-25, 2021, online conference (***oral presentation*** *by Aaron Curtis*).

5. **Yajun Yu**, Salem Werdyani, Megan Carey, Georgia Skardasi, William Pollett, Jane Green, Elizabeth Dicks, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. Novel genetic and clinical prognostic variables in colorectal cancer. *The 33rd Canadian Student Health Research Forum*. August 24-28, 2020, online forum (***poster presentation***).

6. **Yajun Yu**, Megan Carey, William Pollett, Jane Green, Elizabeth Dicks, Patrick Parfrey, Yildiz E. Yilmaz, Sevtap Savas. Predictors of long-term survival outcome risks in a cohort of colorectal cancer patients followed up to 19 years. *The 2019 Canadian Cancer Research Conference (CCRC)*, November 03-05, 2019, Ottawa, Canada (***poster presentation***).

7. Sevtap Savas, Georgia Skardasi, **Yajun Yu**. The SWI/SNF complex subunit genes and their relation to patient survival times in human cancers. *24th World Congress on Advances in Oncology and 24th International Symposium on Molecular Medicine*, October 10-12, 2019, Sparta, Greece (***invited speech*** *delivered by Sevtap Savas*).
8. **Yajun Yu**, Salem Werdyani, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Patrick Parfrey, Yildiz Yilmaz, Sevtap Savas. Copy Number Variations (CNVs) that are associated with the risk of early-relapse in colorectal cancer patients. *The 2016 BHCRI/TFRI (Beatrice Hunter Cancer Research Institute/Terry Fox Research Institute) Cancer Research Conference*, November 07-08, 2016, Halifax, Canada (***poster presentation***).
9. **Yajun Yu**, Salem Werdyani, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Patrick Parfrey, Yildiz Yilmaz, Sevtap Savas. Germline Copy Number Variations (CNVs) that affect genes and relapse-free survival in colorectal cancer. *2016 International Genetic Epidemiology Society (IGES) Meeting*, October 24-26, 2016, Toronto, Canada (***poster presentation***).
10. **Yajun Yu**, Salem Werdyani, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Yildiz Yilmaz, Patrick Parfrey, Sevtap Savas. Structural variants in *TGFBR3*, *STEAP2*, and *FILIP1L* genes may associate with disease outcomes in colorectal cancer. *Target Meeting's 4th World Cancer Online Conference*, May 17-19, 2016 (***oral presentation***).
11. **Yajun Yu**, Salem Werdyani, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Yildiz Yilmaz, Patrick Parfrey, Sevtap Savas. Common Copy Number Variations (CNVs) and disease-free survival in colorectal cancer. *5th Annual Canadian Human and Statistical Genetics Meeting*, April 16-19, 2016, Halifax, Canada (***poster presentation***).

Awards and recognitions

Sep 2020 **A.G. Hatcher Memorial Scholarship**, MUN, St. John's, Canada

Aug 2020	**Selected by MUN Med School to attend the 33rd Canadian Student Health Research Forum**
May 2020	**Human Genetics Graduate Student Award**, Faculty of Medicine, MUN, St. John's, Canada
Oct 2019	**BHCRI (Beatrice Hunter Cancer Research Institute) CRTP (Cancer Research Training Program) featured trainee**
Apr 2019	**The Best Poster Presentation** by a PhD candidate in Genetics Research Forum, Faculty of Medicine, MUN, St. John's, Canada
Mar 2019	**Dr. Angus J. Neary Genetics Scholarship**, MUN, St. John's, Canada
Jan 2019	**BHCRI/TFRI (Beatrice Hunter Cancer Research Institute/Terry Fox Research Institute) Cancer Research Training Program Award**, Halifax, Canada (for 20 months)
Oct 2018	**Faculty of Medicine Dean's Fellowship**, Faculty of Medicine, MUN, St. John's, Canada (for 12 months)
Mar 2018	**Dean's Building a Healthy Tomorrow Award**, Faculty of Medicine, MUN, St. John's, Canada
Mar 2018	**Dr. Roger C. Green Graduate Scholarship in Human Genetics**, Faculty of Medicine, MUN, St. John's, Canada
May 2017	**The Best Graduate Student Presenter Award** for Discipline of Genetics Seminar Series, MUN, St. John's, Canada
Dec 2016	**TPMI (Translational and Personalized Medicine Initiative)/NL SUPPORT Educational Funding Fellowship** (for 24 months), St. John's, Canada
Oct 2016	**Travel Bursary** by BHCRI/TFRI (Beatrice Hunter Cancer Research Institute/Terry Fox Research Institute) to attend their cancer research conference, Halifax, Canada

CHAPTER 1: Introduction

1.1 Incidence and mortality rates of colorectal cancer

Globally, colorectal cancer is the third most common type of cancer and the fourth leading cause of cancer death [1,2]. In 2018, nearly two million individuals on this planet were diagnosed with colorectal cancer and more than one million died of it [1]. The incidence rate of colorectal cancer varies across regions or countries, ranging from 3.3 to 45.3 per 100,000 person-years (age standardized incidence rate; data from GLOBOCAN 2020) [3,4]. Generally, there is a higher incidence rate in developed nations than in developing countries, though such a rate is stabilizing or decreasing in some developed regions (e.g. North America and Europe) and increasing in many developing regions, especially those undergoing a rapid development transition [1,4]. Such a difference might be attributed to different life-styles or environmental exposures (e.g., diet, physical activity) [5,6], or the feasibility regarding diagnosis. Mortality rate of colorectal cancer is also variable (2.3 – 21 per 100,000 person-years; age standardized mortality rate) among different regions, though to a lesser extent compared to the incidence rate [3,4]. However, unlike the incidence rate, mortality rate is generally lower in developed nations compared to developing countries, and the trend of this rate, in recent years, generally goes down in the former ones but up for the latter ones [1,7]. This is possibly due to the success in screening, surveillance, and treatment of colorectal cancer in developed nations and increasing incidence rates in developing countries [8].

In Canada, colorectal cancer ranks the third most commonly diagnosed malignancy, accounting for ~12% (n= ~26300) of total new cancer cases within a year [9]. This disease is also the second leading cause of cancer death in Canada, responsible for ~12% (n= ~9500) of total

cancer deaths [9]. The age-standardized incidence and mortality rates of colorectal cancer in Canada are 31.2 and 9.9 per 100,000 person-years, respectively [3,4]. While the incidence rate is about 11% above the world's incidence rate (19.5 per 100,000 person-years; age-standardized), the mortality rate is close to the world's rate (9.0 per 100,000 person-years; age-standardized) [3,4]. Among all the provinces in Canada, Newfoundland and Labrador (NL) has the highest incidence and mortality rates of colorectal cancer [9]. Canadian Cancer Statistics 2019 [9] projected approximately 600 new cases and around 250 deaths caused by colorectal cancer in 2019 in NL. In this thesis research, studies were mainly performed using the data from colorectal cancer patients in NL, the majority of whom had sporadic colorectal cancer.

1.2 Types of colorectal cancer

Colorectal cancer can be categorized into two different types: hereditary/familial colorectal cancer, and sporadic colorectal cancer [10,11]. Patients with hereditary/familial colorectal cancer (accounts for ~35% of all colorectal cancer cases [12]) usually have a family history of this disease. Among these patients, a portion has known hereditary syndromes (accounts for < 5% of all colorectal cancer cases [12]). Patients with sporadic colorectal cancer are those with no hereditary syndromes or obvious family histories, and they account for the largest part (~65%) of colorectal cancer cases [12]. Both types of colorectal cancer are direct results of the multistep accumulation of mutations/alterations in epithelium cells in the colon or rectum [13], however, their genetic basis may be different.

1.2.1 Hereditary/familial colorectal cancer

Hereditary/familial colorectal cancer accounts for ~35% of total colorectal cancer cases [12,14–16]. Hereditary colorectal cancers usually have known genetic bases and normally have early disease onset (**Table 1.1**). The most well-known hereditary colorectal cancer include Lynch syndrome, familial adenomatous polyposis (FAP), and *MUTYH*-associated polyposis (MAP) [15,17]. Lynch syndrome is the most common type of hereditary colorectal cancer syndrome. It affects 2% to 3% of all colorectal cancer patients [17]. Lynch syndrome is inherited in an autosomal dominant fashion [18]. Genetic basis of Lynch syndrome includes germline mutations in mismatch DNA repair genes, including *MLH1*, *MSH2*, *MSH6*, and *PMS2* [18]. FAP is also inherited as an autosomal dominant disease. It is usually characterized by more than a hundred polyps in the colon and rectum, and around 70% of patients also have extra-intestinal manifestations (e.g., osteomas and dental abnormalities) [19]. FAP accounts for <1% of all colorectal cancers [20] and is caused by germline mutations of the *APC* gene, which is located on 5q22.2. This gene encodes a protein (APC) involved in the WNT singling pathway [21] (**Figure 1.1**). In this pathway, APC binds to β-catenin and thus suppresses the downstream cellular processes, including cell proliferation and survival [21]. When *APC* (or other key genes [e.g., *CTNNB1*, the gene encoding β-catenin] in the pathway) is mutated, this can lead to constant activation of the WNT singling pathway and thus uncontrolled cell proliferation and survival [21] (**Figure 1.1**). MAP is characterized by multiple adenomatous polyps (usually <100 adenomatous polyps, but it can present with hundreds of polyps) in the gastrointestinal tract, and is caused by the germline mutation of *MUTYH*, which is a base excision repair gene on chromosome 1 [22]. Unlike Lynch syndrome and FAP, MAP is inherited in a recessive fashion, with alterations of both copies of *MUTYH* leading to the development of colorectal cancer [23]. Other than these,

germline mutations in some other genes can also lead to hereditary colorectal cancers. Related genes, and their related biological functions/pathways, of some of the hereditary colorectal cancer syndromes are summarized in **Table 1.1**.

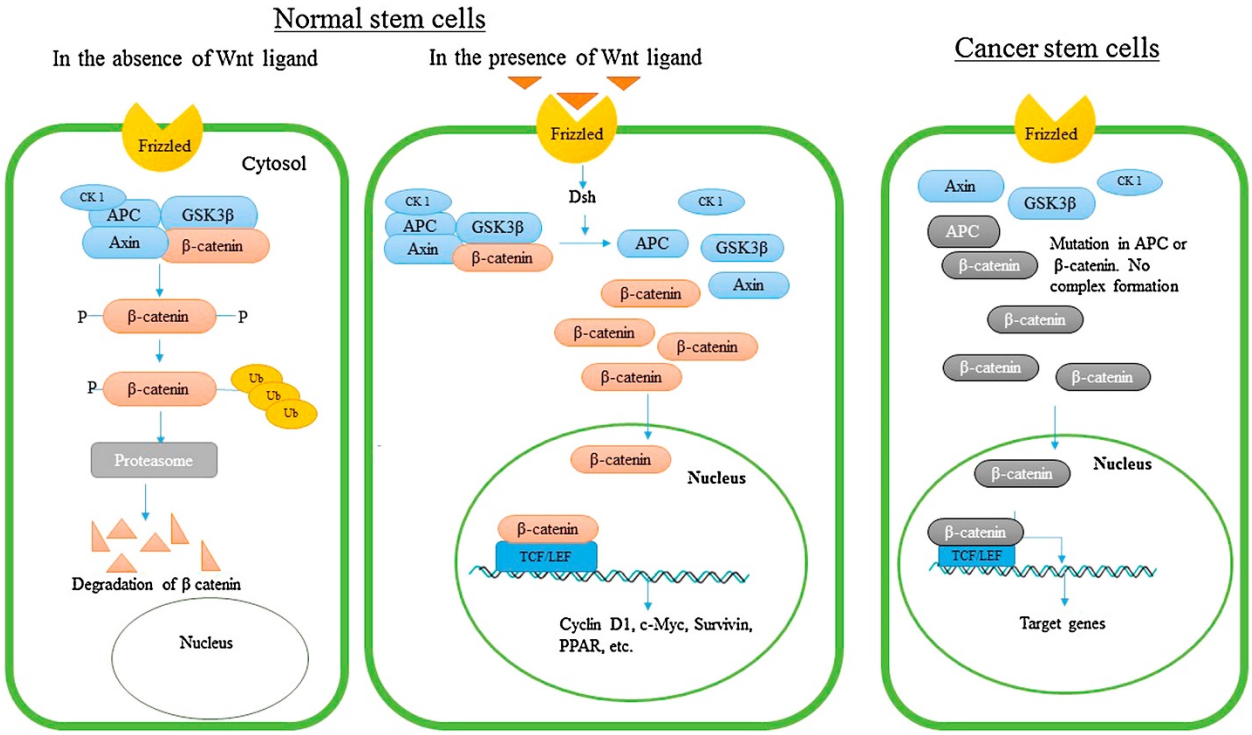


Figure 1.1. The WNT signaling pathway. Reprinted from Vadde et al., 2017 [24] with copyright permission (see **Appendix A**). In normal stem cells, if Wnt ligand is absent, β-catenin would bind to APC and other proteins to mediate its degradation. In such a case, transcription of certain genes involved in cell proliferation and survival would not be activated. Normal stem cells only proliferate when the WNT ligand presents and interrupts the degradation of β-catenin. When APC and other critical genes in the path are mutated, β-catenin will not be degraded (irrespective of the presence of Wnt ligand), leading to uncontrolled constant activation of cell proliferation and survival.

Table 1.1. Genes and age of onset in select known colorectal cancer syndromes. Reprinted from Peters et al., 2015 [25] with copyright permission (see **Appendix A**).

Gene	Hereditary Syndrome	Age of Onset (years)	Pathway/Biological function*
APC	FAP, AFAP	34–43	Wnt signaling pathway
MUTYH	MAP	48–56	Base excision repair
MLH1, MSH2, MSH6, PMS2, EPCAM	Lynch Syndrome	44–56	Mismatch repair
PTEN	Cowden syndrome (includes Bannayan-Ruvalcaba-Riley (BRR) syndrome)	<50 (BRR pediatric onset)	Negative regulator of metabolic signaling
STK11	Peutz-Jeghers Syndrome (PJS)	65	Tumour suppressor responding to changes in cellular energy balance
GREM1,15q13 locus	Hereditary mixed polyposis syndrome (HMPS)	48	TGF-β/BMP signaling pathway
BMPR1A	HMPS, juvenile polyposis syndrome	48, 42	TGF-β/BMP signaling pathway
MADH4/SMAD4	Juvenile polyposis syndrome	42	TGF-β/BMP signaling pathway
POLE, POLD1	Oligopolyposis or Polymerase proofreading-associated polyposis	23–80	DNA repair

*, Many of these pathways interact at multiple levels and as such are not necessarily independent biological mechanisms. Reprinted from Peters et al., 2015 [25] with copyright permission (see Appendix A).

With regard to familial colorectal cancers that do not fit any defined inherited syndromes, such as familial colorectal cancer type X (FCCX) [26,27], their underlying genetic bases still remain unknown [28]. It is assumed that multiple genetic alterations with low or medium penetrance, as well as environmental factors, may together contribute to the development of these familial colorectal cancers [28–30]. Individuals from families with these cancers normally have a higher disease risk compared to the general population but not as high as those with hereditary syndromes [31].

1.2.2 Sporadic colorectal cancer

Sporadic colorectal cancer accounts for the majority (~ 65%) of colorectal cancers [12]. Patients with such a cancer usually have no evident familial history [32] and their ages at disease onset are normally ≥ 50 years [33]. Generally, the development of sporadic colorectal cancer is believed to be contributed to by both the genetic and environmental factors [34–36]. While a number of environmental factors have been known to contribute to the risk of sporadic colorectal cancer, the causal risk variants remain largely unknown.

Though currently little is known about the causal risk variants in sporadic colorectal cancer, we do know that many (somatic) mutations occur in the tumors. Based on the number of somatic mutations acquired and accumulated in tumors, sporadic colorectal cancer can be categorized into two main categories: (1) the hypermutated sporadic colorectal cancer (16%) and (2) the non-hypermutated sporadic colorectal cancer (84%) [35,37]. Mutation rates in hypermutated tumors are higher than 12 per million bases, and that of the non-hypermutated tumors are less than 8.24 per million bases [35,37]. Other than that, the patterns of somatic mutations in these two categories are different (**Figure 1.2**). While mutations of TGF-beta related genes (*ACVR2A* and *TGFBR2*), *APC*, *BRAF*, and mismatch DNA repair (MMR) related genes (*MSH3* and *MSH6*) are more commonly observed in hypermutated tumors, mutations of *APC*, *TP53*, and *KRAS* genes more frequently occur in the non-hypermutated tumors (**Figure 1.2**). *APC*, while it is also included in hereditary/familial colorectal cancer cases, is one of the most mutated genes in all sporadic colorectal cancer cases, irrespective of the category (**Figure 1.2**). With regard to the sequence of occurrence of these mutations (and other molecular characteristics), it can generally be described in three different molecular pathways, which will be reviewed in the next section.

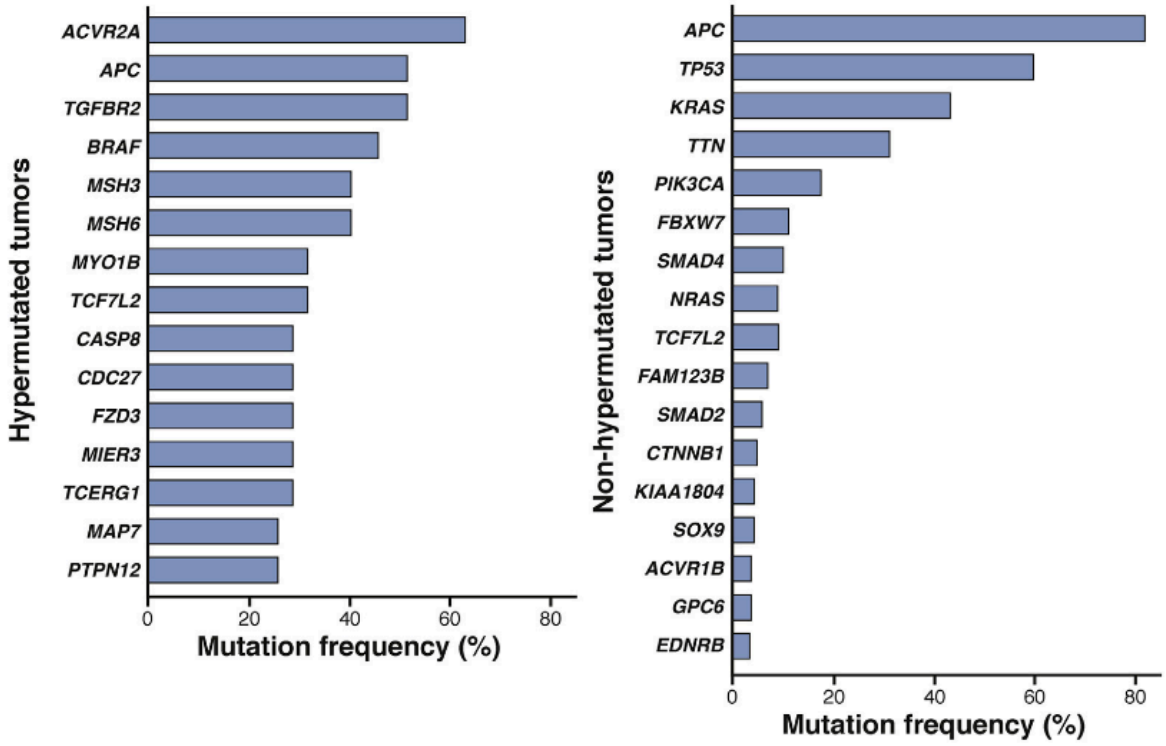


Figure 1.2. Mutation patterns and frequencies of hypermutated and non-hypermutated sporadic colorectal cancers. Reprinted from Carethers & Jung, 2015 [35] with copyright permission (see **Appendix A**).

1.3 Molecular pathways

Normal colorectal cells turn into malignant cells through a process of sequential accumulation of genetic mutations and/or epigenetic alterations [38]. There are three main molecular pathways that describe the progress of normal cells to colorectal tumors: chromosomal instability (CIN), microsatellite instability (MSI), and CpG island methylator phenotype (CIMP) pathways [39] (**Figure 1.3**).

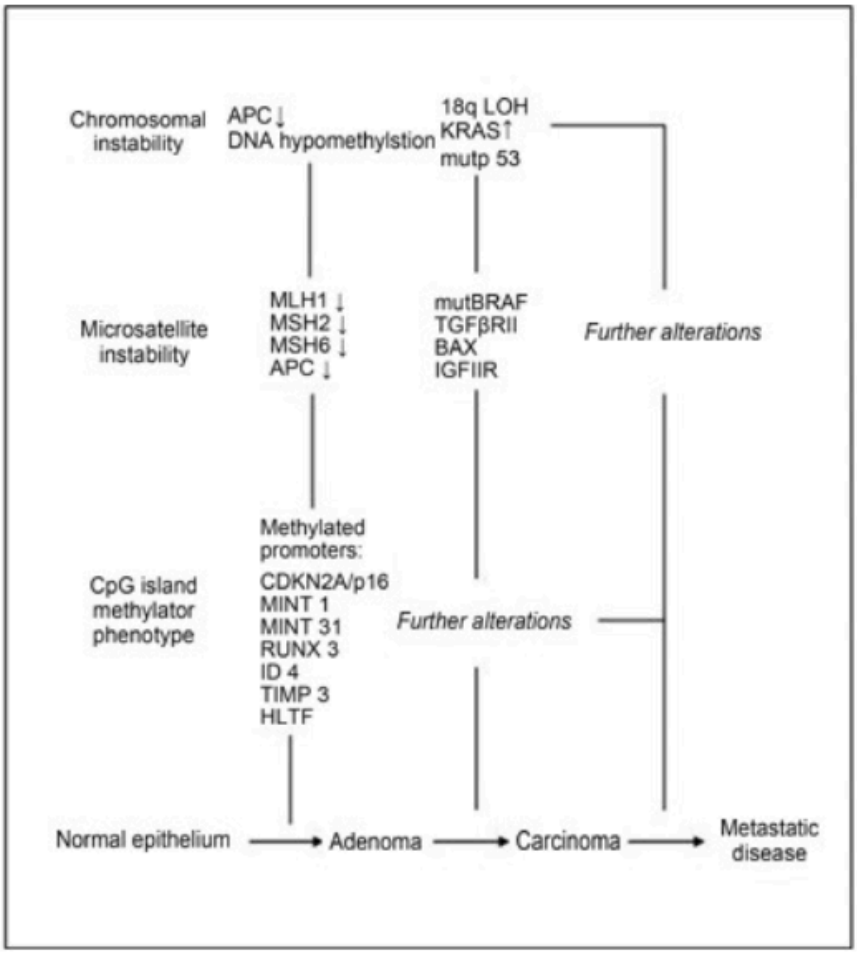


Figure 1.3. Molecular pathways and important related molecular, genetic, and epigenetic changes. Reprinted by permission from Cancer Biology & Medicine (Tariq & Ghias, 2016 [39]). Copyright (2021) by Cancer Biology & Medicine (see **Appendix A**).

1.3.1 Chromosomal instability pathway

CIN pathway, as indicated by its name, is mainly characterised by the chromosomal instability (e.g., structural rearrangements/abnormalities of chromosomes) in tumors. Around 65-70% of total colorectal cancer are due to abnormalities in the CIN pathway [38]. In the transition

process from normal to tumor cells in the colon and rectum, CIN starts with the acquired mutation of the *APC* gene, followed by additional mutations in other genes including oncogenes (e.g., *KRAS*) and tumor suppressor genes (e.g., *TP53*), which are critical for tumorigenesis and drive the malignant transformation [13,39,40]. Accompanied with the accumulation of mutations, CIN related events such as loss or gain of whole chromosomes (i.e., aneuploidy), structural rearrangements/abnormalities of chromosomes (e.g., loss of 18q), and loss of heterozygosity (LOH) occur, resulting in tumors with karyotypic variability among cells [13,39,40]. The exact mechanism of CIN is largely unclear. Possible factors driving CIN include defects in chromosome segregation and DNA replication, and telomere dysfunction [40–42]. With regard to the relationship between CIN and the accumulation of mutations, it is not clear yet which one creates the appropriate environment for the other [31,40].

1.3.2 Microsatellite instability pathway

MSI is found in around 15% of the colorectal tumors [38]. MSI is characterized by the instability of microsatellites, which are nucleotide repeat sequences with motifs of 1-6 base pairs [43]. MSI is defined by a panel of markers (i.e., microsatellites) recommended by the National Cancer Institute [44], and it can be designated as MSI-high (MSI-H) and MSI-low (MSI-L). If there is no such an instability, it is referred to as the microsatellite stable (MSS). While MSS represents no change of markers in the tumor, MSI-H means more than or equal to 30-40% of the markers are unstable, and MSI-L indicates that at least one but less than 30-40% of the markers are unstable [44].

MSI is caused by the alterations of mismatch DNA repair (MMR) genes and hence reflects the defects of mismatch repair system. For familial colorectal cancers with MSI (2%-3% of total colorectal cancer cases), as mentioned in Section 1.2.1, four MMR genes (*MLH1*, *MSH2*, *MSH6*, and *PMS2*) are normally mutated. Sporadic colorectal cancers with MSI (~12% of total colorectal cancer cases [45]), on the other hand, are generally caused by the hypermethylation of *MLH1* promoter and thereof the inactivation of this gene. In the MSI pathway, in addition to mutations or epigenetic alterations of MMR genes, mutations of other critical genes such as *BRAF*, *APC*, *TGFRβII,* and *IGFIIR* promote the development of tumorigenesis. Compared to CIN tumors, MSI-H tumors are usually diploid and have lower frequency of LOH [38].

Colorectal cancer patients with MSI-H usually have better outcomes compared to patients with MSI-L and MSS [46–49], and hence the MSI is usually used as a covariate in survival analysis of colorectal cancer, adjusting the effects of the variables of interest.

1.3.3 CpG island methylator phenotype pathway

CpG island methylator phenotype (CIMP) is characterized by increased methylation of CpG islands across the genome, which are usually enriched in the upstream region (e.g., promoter) of genes, resulting in inactivation of important genes that relate to cancer development and progression (e.g., tumor suppressor genes) [50]. Unlike MSI, CIMP has no standardized panel of loci to define it [50,51]. Many studies use a panel containing the following 5 loci: *MLH1*, *p16*, *MINT1*, *MINT2*, and *MINT31* [52–56]. Based on the degree of methylation, CIMP can be categorized as CIMP-high (CIMP-H; a high degree of methylation) and CIMP-low (CIMP-L; a low degree of methylation) (cut-off values to distinguish CIMP-H from CIMP-L using different

panels can be found in Rhee et al., 2017 [51]), and they occur in around 40% of colorectal cancers (~20% for each) [57]. In the CIMP pathway, global genomic methylation and other mutations/alterations inactivate critical genes (e.g., *BRAF* [mutation of which often occurs in CIMP-H tumors] and *KRAS* [mutation of which often occurs in CIMP-L tumors] [57,58]), and they promote the transition of normal cells to malignant cells [38,50]. Note that the CIMP pathway is not exclusively independent from the other two pathways. For example, promoter hypermethylation of *MLH1* through CIMP pathway can lead to MSI, which is a feature of the MSI pathway [53].

Morphologically, colorectal cancers due to different molecular pathways may follow the same sequence of change from normal cells to tumors. For example, many colorectal tumors due to the CIN and MSI pathways form adenomatous polyps at initial stages of tumorigenesis and then turn into carcinomas (**Figure 1.4**). In fact, the majority of colorectal cancers are adenocarcinomas and follow the adenoma-carcinomas sequence of change [45,59]. Another morphological change in colorectal cancer is via the serrated (or "saw-toothed") neoplasia, which is mainly observed in colorectal cancers due to the CIMP and MSI pathways [60,61].

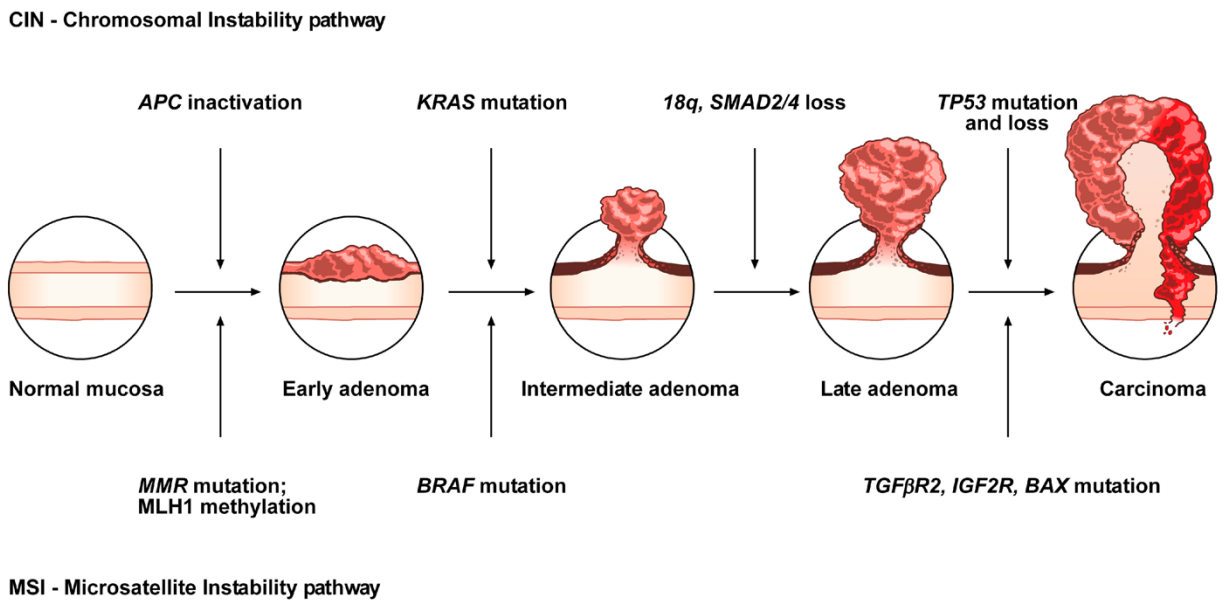


Figure 1.4. Adenoma-carcinoma sequence. Reprinted from De Palma et al., 2019 [61] (This article is under the Creative Commons CC BY license; permission is not required as long as the figure is properly cited).

1.4 Early detection, treatment, and follow-up surveillance

1.4.1 Early detection

Early detection is important in the control of colorectal cancer, as this disease, if it is detected at its early stage, can usually be cured, or patients can achieve better outcomes after treatment. Currently, there are screening programs for early detection of colorectal cancer using different screening approaches.

The main screening approaches for early detection of colorectal cancer include stool-based, imaging, and endoscopic methods. Stool-based methods generally detect the blood in

stool (guaiac fecal occult blood test [gFOBT]; fecal immunochemical test [FIT]) or tumor DNA (stool DNA testing), which indicate possible cancerous lesions in the colon and/or rectum, and they are generally non-invasive, inexpensive, and hence widely used methods for screening of colorectal cancer [62,63]. Image methods of early detection "photograph" the colon and rectum using X-ray, magnetic resonance, or other specific devices/ways [63], and they detect possible adenomatous polyps and/or early-stage tumors in the colon and rectum through imaging [63]. However, these methods (as well as the stool-based methods) may lack the ability of excision and biopsy for further analysis. Luckily, this can be achieved by using the endoscopic methods, such as colonoscopy [62]. Colonoscopy uses a flexible and long endoscope for detecting precancerous and cancerous lesions, and in the meantime, can also perform biopsy and/or polyp removal [64]. This makes colonoscopy a method with high sensitivity and specificity, and thus the current "golden" standard in colorectal cancer screening [63].

1.4.2 Treatment

Treatment of colorectal cancer mainly depends on the disease stage of the patient, considering the potential curability and recurrence of the disease [65,66]. The main modality of treatment in colorectal cancer includes surgery, chemotherapy, and radiotherapy [34,65].

Surgery can be used for curative or palliative purposes in colorectal cancer treatment [65]. For patients with primary tumor only (i.e., stage I-III patients), surgery is considered as a curative therapy, but it is not likely to cure the disease for patients with metastatic tumors (i.e., stage IV patients) [65,67]. For those patients, surgery is usually performed as a means to relieve

symptoms (e.g., pain) and/or achieve better outcomes with the combined use of chemo and radiotherapies [65,67].

With regard to chemotherapy in colorectal cancer, a drug commonly used is the 5-fluorouracil (5-FU) [68]. The main mechanism by which this drug (and its derivatives, e.g., uracil-tegafur [UTF]) exerts its anti-tumor effects is through the inhibition of thymidylate synthase [68,69]. Inhibited thymidylate synthase blocks the synthesis of pyrimidine nucleotides and thus interrupts DNA replication process, leading to death of rapidly dividing tumor cells [68,70]. In addition to 5-FU, two other cytotoxic drugs, irinotecan and oxaliplatin, also act mainly through their impacts on DNA. Irinotecan inhibits topoisomerase I, which is a key component in DNA replication [71]. Oxaliplatin, a platinum derivative, interacts with DNA and prevents its synthesis [72]. These two drugs can be used in combination with 5-FU to obtain better treatment outcomes for patients [65]. In the metastatic setting, the efficacy of treatment may further be increased by using treatments based on monoclonal antibodies, such as bevacizumab, cetuximab, and panitumumab [66,73]. Bevacizumab is an anti-vascular endothelial growth factor (anti-VEGF) antibody, and cetuximab and panitumumab are antibodies targeting epidermal growth factor receptor (EGFR) [66,73]. Both VEGF and EGFR are essential components of important biological pathways involved in tumor cell amplification and migration, thus these monoclonal antibodies can suppress tumor progression by supressing these pathways [73]. Other than the anti-angiogenesis and anti-EGFR agents, in recent years monoclonal antibodies as inhibitors of immune checkpoints (e.g., pembrolizumab and nivolumab) have also been proved to be effective in the treatment of colorectal cancer [73].

Radiotherapy is often used for treatment of locally advanced rectal cancers [65]. As the anatomical confinements of the bony pelvis (in which rectum is located) limit the success of

surgery, a higher rate of positive margins is usually observed in rectal cancer surgery compared to that of colon cancer [65]. In such a case, radiotherapy is utilized as a main method to eliminate residual tumor cells and prevent cancer recurrences [65]. Other than its use after surgery, radiotherapy is also recommended for rectal cancer treatment prior to surgery, and it is believed to be beneficial in reducing tumor burden [74]. Though radiotherapy is mainly performed for rectal cancer, it may also be utilized for locally invasive colon cancer and certain types of metastatic tumors [65,75].

In survival analysis examining the associations of variables of interest with outcomes of colorectal cancer patients, adjuvant chemo- and/or radio-therapies are usually considered as covariates adjusting the effects of variables of interest, as these two treatments are generally believed to be beneficial for the prognosis of colorectal cancer patients and may confound the effects of other variables in the models.

1.4.3 Follow-up surveillance

The main purpose of follow-up for patients after surgery is to detect tumor recurrences at early stages prior to the development of symptoms, and to enable surgical resection of lesions and improve patients' survival in the end [76].

The main modalities of follow-up surveillance include laboratory tests (e.g., carcinoembryonic antigen [CEA]), imaging (e.g., CT scan), endoscopy, and clinical visits [77,78]. Patients may be given different combinations and frequencies of surveillance tests, depending on their disease stages (which surrogate recurrence risk) as well as depending on which guidelines are followed. The guidelines of disease surveillance are shown in **Table 1.2** [79].

With respect to the duration of follow-up, currently there is no consensus [78]. The main guidelines normally recommend 5-years surveillance for colorectal cancer patients (**Table 1.2**), as the majority of local recurrences and distant metastases present within this time-frame following the diagnosis/surgery. However, some studies raised a concern of whether this time-frame should be extended to improve the survival outcomes of patients with late recurrences/relapse, considering the portion of these patients may not be negligible [80–82]. Bouvier and others [80] found that 1 in 12 in men and 1 in 19 in women colon cancer patients (stage I-III) experienced tumor recurrences between 5 and 10 years after diagnosis. In rectal cancer, similar late-recurrence rate (1 in 13 patients) was also observed within the same time-period [82]. Another study showed that among the colorectal cancer patients who experienced disease relapse, around 12% had recurrent tumors after 5 years post-diagnosis [81]. In line with these studies, a study of this thesis research examining prognostic characteristics in colorectal cancer also identified more than 13% of patients with recurrent tumors had their first recurrences/metastases after 5 years follow-up surveillance (Chapter 4).

Table 1.2. Disease surveillance guidelines for colorectal cancer patients. Reprinted from van Der Stok et al., 2017 **[79]** with copyright permission (see **Appendix A**).

Modality	**ASCO 2013** [83]	**ASCRS 2015** [84]	**ESMO (I–III: 2013; IV: 2014)** [85,86]		**#NCCN 2015** [87]		**UK 2011** [88]
	II–III	**I–IV: colon/rectum***	**I–III**	**IV**	**I–III**	**IV**	**I–IV**
History and/or physical exam	Every 3–6 mo. for 5 yrs	• Every 3–6 mo. for 2 yrs • Every 6 mo. in yrs 3–5	• Every 3–6 mo. for 3 yrs • Every 6–12 mo. in yrs 4–5	Every 3–6 mo. for 3 yrs	• Every 3–6 mo. for 2 yrs • Every 6 mo. in yrs 3–5	• Every 3–6 mo. for 2 yrs • Every 6 mo. in yrs 3–5	NA

Serum CEA test	Every 3–6 mo. for 5 yrs	• Every 3–6 mo. for 2 yrs • Every 6 mo. in yrs 3–5	• Every 3–6 mo. for 3 yrs • Every 6–12 mo. in yrs 4–5	Every 3–6 mo. for 3 yrs	• Every 3–6 mo. for 2 yrs • Every 6 mo. in yrs 3–5	• Every 3–6 mo. for 2 yrs • Every 6 mo. in yrs 3–5	Every 3–6 mo. for 3 yrs
Chest CT	Annually, or every 6–12 mo. for high-risk pts, for 3 yrs	Annually for 5 yrs	Every 6–12 mo. for 3 yrs	Every 3–6 mo. for 3 yrs	Annually for 5 yrs	• Every 3–6 mo. for 2 yrs • Every 6–12 mo. in yrs 3–5	Twice over the 3 yrs
Abdominal CT	Annually, or every 6–12 mo. for high-risk pts, for 3 yrs	Annually for 5 yrs	Every 6–12 mo. for 3 yrs	Every 3–6 mo. for 3 yrs	Annually for 5 yrs	• Every 3–6 mo. for 2 yrs • Every 6–12 mo. in yrs 3–5	Twice over the first 3 yrs
Pelvic CT	Annually or every 6–12 mo. for high-risk pts, for up to 3 yrs, or up to 5 yrs for pts with rectal cancer	Annually for 5 yrs	NA	NA	Annually for 5 yrs	• Every 3–6 mo. for 2 yrs • Every 6–12 mo. in yrs 3–5	Twice over the 3 yrs
Liver CEUS	NA	NA	Can substitute for abdominal CT	NA	NA	NA	NA
Colonoscopy	At 1 yr; every 5 yrs thereafter	At 1 yr	At 1 yr; every 3–5 yrs thereafter	NA	At 1 and 4 yrs, then every 5 yrs; annually if advanced adenoma is detected	At 1 and 4 yrs, then every 5 yrs; annually if advanced adenoma is detected	At 1 yr; then every 5 yrs
Proctoscopy (± ERUS)	NA	• Colon cancer: NA • Rectal cancer: Every 6–12 mo. for those with anastomosis, or every 6 mo. after local	NA	NA	NA	NA	NA

		excision, for 3–5 yrs					

ASCO, American Society of Clinical Oncology; ASCRS, American Society of Colon and Rectal Cancer Surgeons; CEA, carcinoembryonic antigen; CEUS, contrast-enhanced ultrasonography; ERUS, endorectal ultrasonography; ESMO, European Society for Medical Oncology; mo., months; NA, not applicable; NCCN, National Comprehensive Cancer Network; pts, patients; yr(s), year(s).

* Same recommendations for colon and rectal cancer, unless noted.

This reference in the original article van Der Stok et al., 2017 [79] provided a website link (https://www.nccn.org/professionals/physician_gls/f_guidelines.asp#site).

1.5 Risk factors for colorectal cancer

Identifying risk factors for colorectal cancer is important, as it can help with the prevention and treatment of this disease. Generally, risk of colorectal cancer is affected by multi-factors. These factors are mainly categorized into two categories: genetic risk factors and environmental risk factors [16]. Though identifying the risk factors of colorectal cancer is not the main aim of this thesis research, as part of a study in this thesis research (Chapter 2), I examined two variants to see if they were associated with the risk/susceptibility of this disease. In the following two Sections (1.5.1 and 1.5.2) of the thesis introduction, the two main categories of risk factors in colorectal cancer are briefly introduced.

1.5.1 Genetic risk factors

As shown in Section 1.2, genetics play a role in colorectal cancer risk. It has been estimated that genetic factors contribute to 12%-13% (based on a family study [89]) or even up to

35% (based on a twin study [16]) of the susceptibility/risk of colorectal cancer. For hereditary colorectal cancers, mutations of certain oncogenes and tumor suppressor genes can eventually result in colorectal cancer. For example, as described earlier, the mutations in the *APC* gene and the DNA mismatch repair genes (e.g., *MLH1*, *MSH2*, *MSH6*, and *PMS6*) can lead to FAP and Lynch syndrome, respectively. While hereditary colorectal cancers is usually caused by a single or a few rare mutations/alterations with large effects (and hence highly-penetrant), other colorectal cancers are normally believed to be affected by joint effects of multiple common genetic variants with small or modest effects as well as environmental factors which are discussed in the following section [25,30,90,91]. In the past fourteen years since the first genome-wide association study (GWAS) on the risk of colorectal cancer was published in 2007 [92], GWASs and related meta-analyses identified 79 loci (with around 100 variants) that were associated with the risk of this disease (**Table 1.3**). Among those identified loci, some (e.g., 8q24.21 [a region that covers *MYC*], 10q25.2 [a region that overlaps with *TCF7L2*], and 12p13.32 [a region that covers *CCND2*]) were identified in multiple studies, strengthening the confidence about their associations with the colorectal cancer risk. However, whether the identified loci/variants are causal loci/variants responsible for tumorigenesis of colorectal cancer remains unclear and warrants further investigation, though some of them were known to be overlapping, within, or near oncogenes (e.g., *MYC*, *CCND2*) or tumor suppressor genes (e.g., *TCF7L2*). Additional information regarding the rs numbers of the identified loci/variants and genes are summarized in **Table 1.3**.

Although a number of risk variants have been reported in colorectal cancer, many other variants contributing to the risk of this disease remain undefined [25,93,94]. Further studies are still needed for identifying additional risk variants. In this thesis research, two INDELs in the *BRM*

gene (an interesting gene, which is involved in chromatin remodeling and has been reported to be associated with the risk of multiple cancers [95–99], but not colorectal cancer prior to my study) were also examined for their associations with disease risk in colorectal cancer (Chapter 2).

Table 1.3. Variants and their loci that have been identified to be associated with the risk of colorectal cancer in GWASs and meta-analyses.

Locus	identified variant(s)	Overlapped or nearby gene(s)	Reference
1p32.3	rs12144319	*TTC22; PCSK9*	94
1p34.3	rs4360494	*FHL3*	94
1p36.12	rs72647484	*CDC42; WNT4*	100
1q25.3	rs10911251	*LAMC1*	101
1q41	rs6691170, rs6687758	*DUSP10*	102
2q24.2	rs448513	*MARCH7; TANC1*	94
2q32.3	rs11903757	*NABP1*	103
2q33.1	rs983402, rs11884596	*SATB2*	94
2q35	rs992157	*PNKD; TMBIM1*	104
3p14.1	rs812481	*LRIG1*	105
3p22.1	rs35360328	*CTNNB1*	105
3q13.2	rs72942485	*BOC*	94
3q22.2	rs10049390	*SLCO2A1*	94
3q26.2	rs10936599	*MYNN*	102
4q22.2	rs1370821	*ATOH1; SMARCAD1*	106
4q24	rs1391441	*TET2*	94
4q31.21	rs11727676	*HHIP*	94
4q32.2	rs35509282	*FSTL5*	107

5p13.1	rs58791712, rs7708610	*PTGER4; LINC00603*	[94,106]
5p15.33	rs2735940, rs78368589	*TERT; CLPTM1L*	[94,106]
5q21.1	rs145364999	*RGMB; CHD1*	[94]
5q31.1	rs647161	*PITX1*	[108]
6p12.1	rs62404968	*BMP5*	[106]
6p21.1	rs4711689	*TFEB*	[109]
6p21.2	rs1321311	*CDKN1A*	[110]
6p21.31	rs6906359	*FKBP5*	[106]
6p21.32	rs9271695, rs2516420	*HLA-DRB1; HLA-DQA1; HLA-B; MICA; MICB; NFKBIL1; TNF*	[94]
6q25.3	rs7758229	*SLC22A3*	[111]
7p13	rs12672022	*MYO1G; SNHG15; CCM2; TBRG4*	[94]
8q23.3	rs16892766, rs2450115, rs6469656	*EIF3H*	[109,112]
8q24.21	rs6983267, rs7014346, rs4313119	*DQ486513; POU5F1; POU5F1P1; DQ515897; MYC*	[92,94,111,113,114]
9p21.3	rs1537372	*ANRIL; CDKN2A; CDKN2B*	[94]
9q22.33	rs34405347	*GALNT12; TGFBR1*	[94]
9q31.3	rs10980628	*LPAR1*	[94]
10p14	rs10795668	*BC031880; LOC389936*	[112]
10q11.23	rs10994860	*A1CF*	[106]
10q22.3	rs704017	*ZMIZ1-AS1*	[115]
10q24.2	rs1035209, rs11190164	*ABCC2/MRP2; SLC25A28; ENTPD7; COX15; CUTC; ABCC2*	[101,105]
10q24.32	rs4919687	*CYP17A1*	[109]
10q25.2	rs12241008, rs11196172, rs10506868,	*VTI1A; TCF7L2; VTI1A*	[109,115,116]
11q12.2	rs174537	*MYRF; FEN1; FADS1; FADS2*	[115]
11q13.4	rs3824999, rs61389091	*POLD3*	[94,110]

11q22.1	rs2186607	*YAP1*	[94]
11q23.1	rs3802842	*LOC120376; FLJ45803; C11orf53; POU2AF1*	[113]
12p13.2	rs2238126	*ETV6*	[117]
12p13.31	rs10849432, rs11064437	*CD9; SPSB2*	[109,115]
12p13.32	rs10774214, rs3217810, rs3217874	*CCND2*	[94,101,108]
12q12	rs11610543	*PRICKLE1; YAF2*	[94]
12q13.12	rs11169552, rs7136702	*DIP2B; ATF1; LARP4; DIP2B*	[102]
12q13.3	rs4759277	*STAT6; LRP1; NAB2*	[94]
12q24.12	rs3184504	*SH2B3*	[105]
12q24.21	rs72013726	*MED13L*	[106]
12q24.22	rs73208120	*NOS1*	[105]
13q13.3	rs7333607	*SMAD9*	[94]
13q22.1	rs78341008	*KLF5*	[94]
13q34	rs8000189	*COL4A2; COL4A1; RAB20*	[94]
14q22.2	rs4444235	*BMP4*	[118]
14q23.1	rs17094983, rs17094983	*RTN1; DACT1*	[94,119]
15q13.3	rs4779584, rs17816465	*SCG5; GREM1*	[94,112]
15q22.33	rs56324967	*SMAD3*	[94]
16p13.2	rs79900961	*C16orf72*	[100]
16q22.1	rs9929218	*CDH1*	[118]
16q23.2	rs9930005	*MAF*	[94]
16q24.1	rs847208, rs2696839	*FENDRR; LOC146513*	[106,120]
17p12	rs1078643	*LINC00675*	[94]
17p13.3	rs12603526	*NXN*	[115]
17q24.3	rs983318	*LINC00673*	[94]

17q25.3	rs75954926	*RAB40B; METRLN*	[94]
18q21.1	rs4939827, rs12953717, rs4464148	*SMAD7*	[121,122]
19p13.11	rs34797592	*KLF2*	[94]
19q13.11	rs10411210, rs7252505	*RHPN2; GPATCH1*	[118,123]
19q13.2	rs1800469, rs2241714	*TGFB1; B9D2*	[115]
19q13.32	rs56848936	*SYMPK*	[123]
19q13.43	rs73068325	*TRIM28*	[94]
20p12.3	rs961253, rs2423279, rs28488, rs994308	*HAO1; BMP2*	[94,108,118]
20q13.12	rs6065668, rs6031311	*TOX2; HNF4A*	[94,120]
20q13.13	rs6066825, rs1810502	*PREX1; PTPN1*	[105,106]
20q13.33	rs4925386, rs6061231, rs2738783	*LAMA5; RPS21; TNFRSF6B; RTEL1*	[94,102,109]
Xp22.2	rs5934683	*SHROOM2*	[110]

1.5.2 Environmental risk factors

A number of environmental factors are linked to the risk of colorectal cancer. These include the factors that relate to diet, physical activity, obesity, alcohol consumption, cigarette smoking, and others [124].

It has been reported that diet rich in fat, especially animal fat, can elevate the risk of developing colorectal cancer, possibly through their stimulation to produce bile acids [125–127]. With the involvement of the bacteria flora in the colon, these acids dehydrogenate to form deoxycholic and lithocholic acids, which can promote carcinogenesis in colon and rectum [125–127]. Low fiber, high calorie diets, and red meat consumption are other diet-related factors that link to

increased risk of this disease [124]. The exact reason why they confer a higher risk maybe complex and remains poorly understood. However, it is possibly through forming toxic substances (e.g., N-nitroso compounds) and influencing insulin sensitivity and microbial composition and metabolism in guts [128,129]. Other than these, low physical activity (e.g., sedentary behavior) and obesity also contribute to the risk of colorectal cancer [124]. These environmental/lifestyle factors usually are collectively termed as the "western lifestyle", which is regarded as one possible contributor to the increased incidence rates of colorectal cancer in countries/regions under rapid economic growth, or populations migrated to western countries [8,124,130,131].

Alcohol consumption and cigarette smoking are two other known risk factors for colorectal cancer [132]. The effect of alcohol on the risk of colorectal cancer may be attributed to its impact on the synthesis of folate, which is a critical component involved in DNA synthesis and repair [133]. Alcohol can be metabolized into acetaldehyde which degrades folate [126,134], leading to impaired DNA and chromosomes, and thus contributing to carcinogenesis [133]. With regard to cigarette smoking, its effect on the risk of colorectal cancer is attributed to the cigarette carcinogens, spreading to the colon and rectum through gastrointestinal tract or blood vessels [126,135].

Other than the modifiable environmental risk factors mentioned above, non-modifiable factors such as age, sex, and inflammatory bowel disease (IBD) also influence the risk of colorectal cancer [132,136]. Generally, men and individuals with older age or having IBD have an increased risk of colorectal cancer [132,136].

1.6 Prognostic markers in colorectal cancer

Other than predicting disease risk/susceptibility, predicting the risk of outcome events (e.g., death, local recurrence, or metastasis) is also an important aspect in research and clinic management of colorectal cancer patients. Investigating prognostic markers that can predict the risk of outcome events in colorectal cancer patients is one of the main aims of this thesis research. Prognostic markers have a prediction ability because they are associated with the survival outcomes of colorectal cancer patients. An important clinical utility of these markers is that they can be used to guide patients' treatment and management strategies to improve their survival times. For example, a patient with a high risk of tumor recurrence can be surveilled more intensively in his/her follow-up, making any returned disease to be detected at the earliest stage and treated accordingly.

There are studies that have been performed to identify prognostic markers from various factors, including clinico-demographic factors, biomarkers, and genetic variants. While many of these markers are not integrated into clinical management of patients, it's nevertheless an important clinical research area for future utility.

1.6.1 Factors examined as prognostic markers

1.6.1.1 Clinico-demographic factors

A number of clinico-demographic factors can predict the prognosis of colorectal cancer patients. Disease stage is the most well-known established prognostic factor in this disease [137,138]. Generally, a higher disease stage is associated with worse patient outcomes. There are different

staging systems being used in the clinical practice, and currently the most commonly used and recommended one is the TNM classification of the American Joint Committee on Cancer/International Union Against Cancer (AJCC/UICC) [139,140]. The TNM staging system classifies tumors into four main groups (stage I, II, III, and IV) and a number of sub-groups by summarizing the information of tumor invasiveness (T stage), the number of local-regional lymph nodes with tumor cells (N stage), and whether metastatic disease (M stage) is present [140]. TNM staging system has a significant prognostic value in clinical practice, which helps guide the treatment and follow-up management of patients [140]. Other clinico-demographic prognostic factors include age, tumor location, tumor grade, and tumor budding which is defined as small clusters of tumor cells at the invasive front of tumors [138,141–145]. Older patients usually have a poor prognosis, though it has been reported that patients with age < 40 or 50 tend to present with later disease stages and have more aggressive disease features [146,147]. With regard to tumor location and grade, rectal and poorly differentiated tumors are more aggressive compared to colon and well/moderately differentiated tumors, and patients with these tumors usually have a shorter survival time [142,143,148]. Tumor budding is also regarded as an adverse prognostic marker in colorectal cancer, with patients presenting this characteristic in their tumors usually have worse outcomes [144].

1.6.1.2 Biomarkers

Other than the clinico-demographic features, biomarkers (or molecular markers) in tissue or blood may also be prognostic markers in colorectal cancer. A well-known biomarker in colorectal cancer is the MSI status which can be used to predict the efficacy of 5-FU based chemotherapy (to be specific, MSI-H predicts an anti-5-FU effect) [149]. In colorectal cancer,

patients with MSI-H normally have better outcomes compared to patients with MSS/MSI-L, possibly due to the fact that MSI-H tumors are often characterized by immune infiltration, which suppresses tumor metastasis [150]. Other biomarkers such as tumor *BRAF* and *KRAS* mutations (which indicate poor response of tumors to anti-EGFR target therapies) were also reported to be associated with worse outcomes of colorectal cancer patients [151,152]. *BRAF* and *KRAS* are the key components of the RAS/RAF/MAPK pathway, and mutations in these genes can lead to aberrant activation of the pathway, promoting tumor progression and migration [153,154]. Patients with tumors containing these mutations hence have worse outcomes compared to patients with no such mutations in their tumors [155–157], although some studies reported no associations between *KRAS* mutations and the survival of patients [158,159]. In addition to these biomarkers, others such as 18q LOH [138], carcinoembryonic antigen [160,161] and microRNAs (e.g., microRNA-21) [162,163] were also widely investigated and reported to have possible prognostic value in colorectal cancer.

1.6.1.3 Genetic variants

Although some genetic mutations (e.g., somatic *BRAF* and *KRAS* mutations in tumors) are already known to have prognostic value, the majority of genetic variants (especially germline variants) are still poorly understood for their potential as prognostic markers in colorectal cancer. Among the studies examining the associations between variants and survival outcomes of colorectal cancer patients, most of them are candidate variant/gene/pathway studies. Such studies focus on a small number of variants in the human genome, and they usually require prior knowledge of the examined variants or their related genes/pathways that may implicate prognostic effects. For example, the 5-FU metabolism pathway is known to be important in 5-FU

efficacy in colorectal cancer treatment, thus variants in genes of this pathway were of great interest and some were reported to be associated with outcomes of patients [164–166]. A number of variants that were previously identified as disease-risk loci (e.g., rs9929218 in *CDH1* [118]) were also of interest and some were found to have prognostic associations in colorectal cancer [167–174]. Although a small portion of variants from candidate variant/gene/pathway studies were replicated for their associations with prognosis in colorectal cancer [175], further validations are still needed before they are widely used in clinic.

Unlike candidate variant/gene/pathway studies, GWASs examine a large number of variants in the genome and do not require prior knowledge of the examined variants and related genes/pathways. Because of these advantages, GWASs are performed to identify variants that can potentially be prognostic markers in different diseases. In colorectal cancer, currently, only a few GWASs were performed to examine the associations between genetic variants and disease outcomes prior to my research, and they identified a limited number of variants having prognostic associations (**Table 1.4**). The first GWAS that was performed in 2015 [176] identified promising SNPs, but it detected no genome-wide significance level associations (p-value < 5×10^{-8}) with overall survival (OS) and disease-free survival (DFS) in patients with non-MSI-H, colorectal, colon, or rectal tumors. Another study [177] focused on stage I-III patients and identified ten SNPs (of which seven were located in introns of *EPHB1*, *FHIT*, and *MIR7515*) that had associations with metastasis-free survival. Pander et al. [178] examined a cohort of metastatic patients, who were treated with combination chemotherapy as part of the CAIRO-2 trial. These authors identified a SNP in an intron of *GnT-IVa* that was associated with progression-free survival (PFS). In another study [179], no significant associations were detected when all patients were analyzed with respect to OS and disease-specific survival (DSS). However, in patients with

stage IV disease, two SNPs near the *ELOVL5* gene were associated with OS, and one of them was also associated with DSS. In a recent study, Penney et al. [180] examined associations in stage II and III colon cancer patients, who were recruited to two clinical trials. As a result, two highly linked SNPs near the *SKAP2* gene were found to be associated with OS, but not DFS. Another recent study performed by Innocenti et al. focused on stage IV colorectal cancer patients, and a SNP in *AXIN1* was identified to have an association with OS [181]. Rs numbers of identified variants, their related genes, and outcomes in these GWASs are summarized in **Table 1.4**. Obviously, additional studies (including GWASs and candidate variant/gene/pathway studies) in identifying potential prognostic variants in colorectal cancer are needed, and this is a major part of my work in this thesis research (Chapters 2, 3, and 5).

Table 1.4. Genome-wide association studies (GWASs) on prognosis in colorectal cancer and the identified variants that were associated with survival outcomes.

Variant	Gene	Outcome	Reference
rs11644916	*AXIN1*	OS	[181]
rs5749032, rs2327990, rs11918092, rs3732568, rs2366964, rs1563948, rs11694697, rs11692570, rs2219613, rs1145724	*EPHB1*, *FHIT*, and *MIR7515* *(seven variants are located in genes)*	MFS	[177]
rs76766811	*SKAP2*	OS	[180]
None	None	DFS	
rs209489	*ELOVL5*	OS	[179]
rs17544464, rs209489	*ELOVL5*	DSS	
rs885036	*GnT-Iva*	PFS	[178]
None	None	OS, DFS	[176]

DFS, disease-free survival; DSS, disease-specific survival; MFS, metastasis-free survival; OS, overall survival; PFS, progression-free survival.

1.6.2 Two types of prognostic markers regarding their associations over time

Other than categorizing prognostic markers based on the nature of variables (e.g., clinico-demographic variables and genetic variants), they can also be categorized into sub-types based on their associations with outcomes over time. Depending on their pattern of associations over time, prognostic markers can be categorized into two types: (1) those with constant associations; and (2) those with non-constant associations (i.e., time-varying associations). My research included examining these types of markers (Chapter 2-5).

1.6.2.1 Prognostic markers with constant associations

These types of prognostic markers are markers that have constant associations with disease outcomes over time (i.e., from the time of disease diagnosis/surgery to the time of event or end of follow up). At any time point after diagnosis, patients in one group of such a marker have the same higher level of outcome-risk compared to patients in the other group of the marker. This time-independent "constant association" feature is important, as prognostic markers with this feature can tell that a group of patients have a constant higher/lower outcome risk compared to the other patients during the time-period from disease diagnosis/surgery to the event of interest or end of follow up.

Many studies reported prognostic associations. However, considering the fact that many of them did not check the possible non-constant associations (for more details, see Section 1.7.1.1) [182,183], a portion of these factors may be misclassified as factors with constant

associations, or they may miss markers with time-varying associations. This is a limitation of many reported studies [184,185]. Factors that were reported to have time-varying associations with outcomes are discussed in the following section.

1.6.2.2 Prognostic markers with time-varying associations

While many prognostic markers have constant associations with outcomes, others may have non-constant associations over time. These markers thus are called prognostic markers with time-varying associations. For such markers, their associations with outcomes may appear, diminish, become stronger or weaker, or even change their directions (e.g., from protective to detrimental, or vice versa) over time [186–188]. Because of that, they can be very useful in predicting outcome risks of patients during specific time-periods of follow-up. For example, a prognostic marker that associated with the risk of recurrence within the first few years post-diagnosis can predict recurrent-disease risk within the initial years of follow-up (early-outcome marker). Likewise, a prognostic marker that has an association with the recurrence after a certain number of years post-diagnosis can predict the outcome risk in those years onward (late-outcome marker). Such information is critical to the personalized disease-management as patients can be surveilled and treated in compliance with their outcome-risk patterns predicted by the time-varying prognostic markers.

Currently, identifying prognostic markers with time-varying associations is not a widely investigated research field in colorectal cancer. Prior to my research, there were only eight studies that investigated time-varying associations of clinico-demographic and genetic factors with outcomes in colorectal cancer, and they identified a small number (n = 9) of factors that

may have such associations (**Table 1.5**). Seven of the identified factors were clinico-demographic factors and only two were genetic factors (**Table 1.5**), indicating more studies need to be performed in this field, especially on genetic factors. The findings of time-varying associations in the eight studies are summarized below.

In 1996, Roncucci and co-authors [190] found that tumor location had a possible time-varying association with disease-specific survival (DSS) in colorectal cancer. Within the first 2 years post-diagnosis, patients with rectal cancer seem to have a lower risk of death compared to patients with colon cancer, but the direction of this risk was reversed after that [190].

Later on between 1999-2003, three studies [191–193] reported more clinico-demographic factors with non-constant associations by investigating a colon cancer dataset from France. Among three studies, the study performed by Bolard and others [191] investigated the tumor site (right vs left) for its effect/association with DSS in colon cancer, and they found that its effect/association was changing over time, from detrimental effect to protective effect. This study also found that age at diagnosis, disease stage, and period of diagnosis had time-varying effects/associations with DSS. Compared to patients with age at diagnosis <65, disease stage I, and periods of diagnosis later than year 1978, patients with age at diagnosis >75, disease stage III, IIIb, or IV, and the period of diagnosis between 1976 and 1978 had worse outcomes within the first few months/years post-diagnosis, but better outcomes after that. Giorgi and others [192] also investigated variables with time-varying effects/associations on DSS (using a different modeling method). Age at diagnosis was again identified to have time-varying associations with DSS, with elderly patients having higher risk of mortality compared to younger patients within the first 6 months, but this association diminishing after that. Other than that, period of diagnosis was also identified to have non-constant association. For example, patients with a diagnosis

between 1985 and 1987 had an early protective effect but a detrimental effect later on, compared to patients diagnosed between 1976 and 1978. Regarding other periods of diagnosis and disease stage, they also seem to have changed effects over time, however, their association patterns were not specified in the study. Unlike the two studies just mentioned which focused on DSS, the third study [193] using the colon cancer dataset from France was performed on overall survival (OS). This study found that tumor site (right vs left) had a possible time-varying association in the univariate analysis. Patients with their tumors in the right colon had an increased risk of death compared to patients with their tumors in the left colon within the first-year post-diagnosis, but not after that. They also compared older patients with younger patients for their OS, and they found that age at diagnosis was another factor with time-varying associations with OS. High risk of death was observed for older patients only within the first year after diagnosis, but not in the second and following years. Disease stages III and IV (stages were classified based on the Dukes staging system; see the legends of **Table 1.5**) were also found to have time-varying associations compared to stage I. While stage III patients had an increased and then a decreased risk of death during the follow up, stage IV patients had a fluctuating risk pattern over time (the risk was low in the first month, and then it increased till the year 2 post-diagnosis. After that, the risk decreased). Other than these factors, period of diagnosis was detected to have possible time-varying associations with OS as well. For example, patients with a diagnosis between 1982 and 1986 had better outcomes compared to patients with a diagnosis between 1977 and 1981 immediately after their surgery, but later on outcomes of these two patient groups had little difference. A similar association pattern was found for patients with a diagnosis between 1986 and 1991.

Another paper published in 2003 by Zahl [194] suggested that age (above 70 years), regional cancer, and tumor location/site (pelvic colon) may have time-varying associations with mortality. The risk of mortality of patients with age above 70 was increased within the first-year post-diagnosis, then there was no difference of this risk in the following 3 years; after that, the risk was increased again. Regarding regional cancer, its association pattern was no association at first, then an increased risk appeared, followed by a lower risk after 2 years post-diagnosis. Similarly, the association of tumor location/site (pelvic/sigmoid colon vs ascending colon) appeared after ~2 years post-diagnosis, but not before that.

A more recent study (done by Liu and others in 2017) [195] reported that sex and tumor grade may have possible time-varying associations with overall survival (OS) in colon cancer. In this study, female patients were found to have better outcomes compared to male patients, and the "protective effect" of being a female became stronger over time. Grade III or IV tumors, on the other hand, were associated with an increased risk of death (or had a "detrimental effect"), and this association also became stronger over time. Other factors such as tumor location/site and disease stage were also identified to have potential time-varying associations with OS in this study.

There are two additional studies that investigated genetic variants for their potential time-varying associations in colorectal cancer (**Table 1.5**). The first one (performed by Pavelitz and others) [196] examined a somatic mutation of the *MRE11* gene in stage III colon cancer patients, and it showed that patients with this mutation had worse overall and disease-specific survival (OS and DFS) within the first ~3.5 years post-diagnosis, but after that, they had improved outcomes compared to patients without this mutation. The other study [177] was published in 2019 by Penney and others, and identified a common germline SNP between *CECR2* and *CECR3* with

a potential time-varying association with metastasis-free survival (MFS) in stage I-III colorectal cancer patients. Though the study investigated the SNP in the mixture cure model (see Section 1.7.2.7) which may not give a specific cut-off time point for the change of the associations, a change point of around 2 years was observed from the Kaplan-Meier curve of this SNP, showing that patients with GG genotype had worse outcomes compared to patients with AA or AG genotypes within the first ~2 years post-diagnosis, but after that they had better outcomes.

With regard to the outcomes examined in these studies, most of the studies focused on OS and DSS. Outcomes other than OS and DSS have not been widely investigated. As shown in **Table 1.5**, only two studies investigated MFS or DFS. Hence, time-varying associations of factors with different outcomes (e.g., recurrence-free survival, RFS; recurrence/metastasis-free survival, RMFS) are also worth to be investigated, as I have done in Chapters 4 and 5. In genetics-related studies on time-varying associations, DSS and RMFS are also the outcomes that were not examined prior to my research (Chapter 5).

Regarding the mechanisms of time-varying associations, in general, they are not quite clear yet. However, for some clinico-demographic factors, plausible mechanisms were proposed. For example, the time-varying associations of disease stage may be attributed to the complications of surgery, which leads to a higher risk of death for advanced stage patients in the early years after diagnosis (but later on, this risk decreases) [197,198]. Another example is age at diagnosis. The changed risks of death (of elderly patients) over time may be caused by the aging of the study population [191,193]. While some clinico-demographic factors have plausible explanations for their time-varying associations, as of today, most of the clinico-demographic variables and genetic variants have no explained mechanisms for their non-constant associations

over time. Further investigation on the mechanisms of time-varying associations are therefore warranted and can yield interesting knowledge.

Table 1.5. Factors reported to have potential time-varying associations with outcomes in colorectal cancer.

Type of variable	Variable		Outcome	Reference
Clinico-demographic factor	Tumor location/site	Rectum vs. colon	DSS	(Roncucci et al., 1996) [190]
		Right vs. left	OS	(Quantin et al., 1999) [193]
		Right vs. left	DSS	(Bolard et al., 2001) [191]
		Left vs. right	OS	(Liu et al., 2017) [195]
		Pelvic colon vs. ascending colon	Did not clearly specify, but seems to be DSS	(Zahl, 2003) [194]
	Age at diagnosis	≥65 vs. <65	OS	(Quantin et al., 1999) [193]
		65–74 vs. <65 ≥75 vs. <65	DSS	(Bolard et al., 2001) [191]
		65-74 vs. ≤64	DSS	(Giorgi et al., 2003) [192]
		>70 vs. 20-69	Did not clearly specify, but seems to be DSS	(Zahl, 2003) [194]
	Disease stage	III vs. I [#] IV vs. I [#]	OS	(Quantin et al., 1999) [193]
		III vs. I IIIb vs. I IV vs. I	DSS	(Bolard et al., 2001) [191]

		II vs. I III vs. I IIIb vs. I IV vs. I	DSS	(Giorgi et al., 2003) [192]
		III vs. I/0 IV vs. I/0	OS	(Liu et al., 2017) [195]
	Period of diagnosis (year)	1982-1986 vs. 1977-1981 1986-1991 vs. 1977-1981	OS	(Quantin et al., 1999) [193]
		1979–1981 vs. 1976-1978 1982–1984 vs. 1976-1978 1985–1987 vs. 1976-1978 1988–1990 vs. 1976-1978	DSS	(Bolard et al., 2001) [191]
		1979–1981 vs. 1976–1978 1982–1984 vs. 1976–1978 1985–1987 vs. 1976–1978 1988–1990 vs. 1976–1978	DSS	(Giorgi et al., 2003) [192]
	Regional cancer	Regional cancer vs. localized cancer	Did not clearly specify, but seems to be DSS	(Zahl, 2003) [194]
	Sex	Female vs. male	OS	(Liu et al., 2017) [195]
	Tumor grade	III vs. I IV vs. I	OS	(Liu et al., 2017) [195]
Genetic variant	Somatic *MRE11* mutation	Mutation vs. no mutation	OS, DFS	(Pavelitz et al., 2014) [196]
	rs5749032	GG vs. AA + AG	MFS	(Penney et al., 2019) [177]

[#], I, Dukes A tumors; III, Dukes C resected for potential cure, designated as "C curative"; IV, Dukes C with palliative treatment, metastatic, or unclassified tumors (so-called U) designated as C palliative, D, U. (from Quantin et al., 1999 [193]).

1.7 Detecting and modeling time-varying associations

Although time-varying associations in prognosis are important and have been investigated in some studies, many survival studies did not check possible time-varying associations for examined variables, and just assumed that those variables were factors with constant associations over time [182,183]. Because of that, inappropriate methods may be used for modeling, especially when the examined factors were actually factors with non-constant associations. If so, time-varying associations can be entirely missed in these analyses, and as a result, the effects/associations can be over- or under-estimated and the findings may be misleading [185]. This highlights the importance of checking the possible time-varying associations and examining such associations with appropriate models.

1.7.1 Detecting time-varying associations

Time-varying associations can be detected by taking advantage of the checking of an underlying assumption of the Cox proportional hazards (PH) regression model. This model is the most widely used statistical method in survival analysis when analyzing medical and biomedical data [199,200]. The hazard function under the Cox PH model can be written as:

$$h(t|x) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \dots \beta_m x_m),$$

where $h(t|x)$ is the hazard function, which measures patients' risk of experiencing the event of interest at time t (i.e., instantaneous risk), given the patients have survived till time t and given the covariates $x = (x_1, x_2, \dots, x_m)$. $h_0(t)$ is the baseline hazard function, which is unspecified but estimated from the study data, and it describes the instantaneous risk for patients with only

baseline levels of all variables (i.e., $X_1=X_2=X_3\ldots=X_m=0$). β_i is the coefficient for the variable X_i (i = 1, 2, …, m).

The measure of effect or association used in the Cox PH model is the hazard ratio (i.e., HR). The HR is the ratio of hazards between patients in a specific group (or with a specific value) and patients in the reference group (or with a reference value). Mathematically, the HR that corresponds to a unit increase in variable X_i (and holding other variables fixed; in other words, with no change of groups/values in other variables) can be written as:

$$HR = h(t)_{Xi = a+1} / h(t)_{Xi = a} = \exp(\beta_i),$$

where "a" represents the reference group/value of variable X_i. Because the $\exp(\beta_i)$ is a constant (not a function of time), thus the HR of variable X_i is also a constant independent of time. This also means that hazards of patients in the compared two groups of X_i are proportional over time (i.e., proportional hazards).

1.7.1.1 Proportional hazards assumption

Proportional hazards (PH) assumption is a key assumption of the Cox PH model [199,200]. This assumption means that the hazards of different categories or levels of a factor remain proportional (i.e., the hazards ratio [HR] stays the same) over time. In other words, the association between the factor of interest and the survival outcome remains constant. When the PH assumption is violated, it suggests that the HR does not always remain the same, and either the magnitude or direction of the association changes during the interval of follow-up (i.e., time-varying association) [185,201]. In such a case, the classical Cox PH model cannot be applied for analysis. However, by taking advantage of the violation of the PH assumption, we can identify

the factors that have possible time-varying associations with outcomes, as I have done in my research projects. Note that variables with time-varying associations include early-outcome or late-outcome markers, depending on the characteristics of their associations over time.

Checking the PH assumption of Cox PH models is not a common practice in survival analyses in cancer research. This has been shown by several studies. In 1995, Altman and others [184] reviewed papers with survival analyses from five clinical oncology journals, and they found that only two papers (5% of all papers that used the Cox models) checked the PH assumption. Another study [183] that published in 2019 checked survival-related clinical trials between 1995 and 2014, and it showed that only 11% explicitly reported the PH assumption test results. A recent review study (in 2020) [182] on cancer documented 28% of the survival studies between 2012 and 2018 using the Cox modeling checked the PH assumption. Though this rate is already much higher than that of Altman's study published in 1995, there is still a large portion (72%) of survival studies in cancer did not check or report the PH assumption. Because of that, there might be many studies that mis-used the analytic models when assuming proportional hazards for examined factors, and this may result in unreliable inference and missing discovery of factors with time-varying associations, and thus the potential early- or late-outcome markers. Considering these, all survival analyses using Cox PH models, ideally, should check the PH assumption not to miss any significant time-varying associations, to avoid miss-detection of associations, and thus to obtain reliable results. It is of note that checking the PH assumption in survival analysis has been followed throughout the studies of the current thesis research (Chapters 2-5). In particular, a study (Chapter 3) of this thesis research directly showed that two variants with time-varying associations cannot be detected in models under the setting of

assuming the PH assumption is satisfied for all variants examined, further underscoring the importance of checking the PH assumption in identifying factors with time-varying associations.

1.7.1.2 Methods of checking the PH assumption

The main methods to test the PH assumption include the graphic and analytic methods.

The graphic methods mainly apply to categorical factors [185,202], and include several methods. A simple method to assess the PH assumption is to plot the Kaplan-Meier (KM) curves [185]. In these plots, the curves represent different groups/levels of the factor of interest, and the separation of these curves over time suggests a satisfaction of the PH assumption. Such assumption is violated when the curves coming toward each other (or even crossover) or drifting apart at a changed rate over time. Another commonly used graphic methods is to plot the log(-log(S(t))) function, where S(t) is the survival function [185,203]. When the PH assumption is satisfied, the curves (which represent different groups/levels of the factor) are approximately in parallel (**Figure 1.5A**). Otherwise, the assumption is violated (**Figure 1.5B**). Generating residual-based plots is another graphic way to check the PH assumption [203]. For example, in plots with scaled Schoenfeld residual-based coefficients against time, the PH assumption is satisfied when horizontal lines (indicating constant HRs) are observed [204]. Inclined or fluctuating lines suggest violation of the assumption (**Figure 1.6**). The advantage of graphic methods is that they can visually show whether the PH assumption is held or not. However, decisions based on these graphics can be subjective. In practice, sometimes it can be hard to tell if a factor violates the PH assumption or not using graphical methods, because the boundary between satisfaction and

violation of the assumption may not be very clear, especially when the examined factor has only a subtle time-varying association.

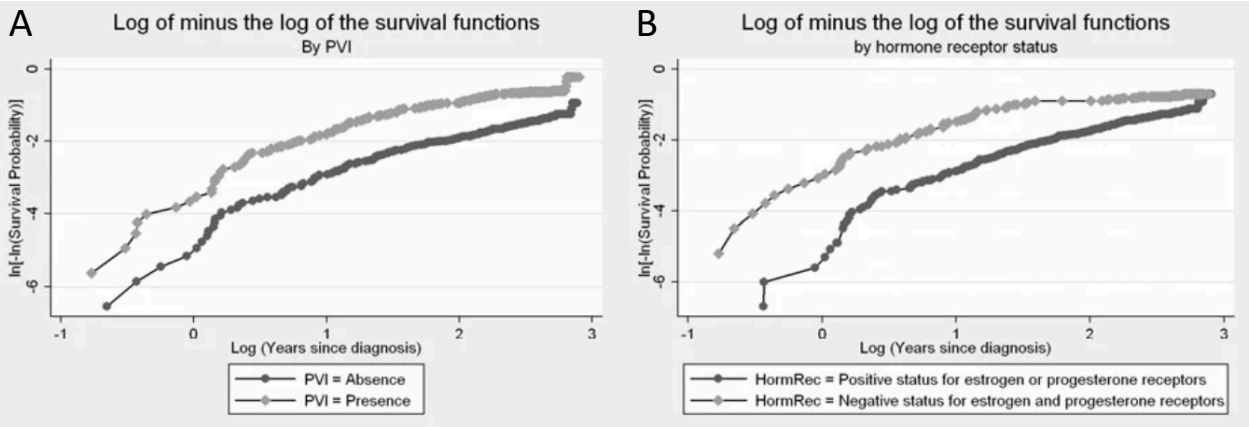


Figure 1.5. Log(-log(S(t))) plot for checking the PH assumption. Reprinted from Bellera et al., 2010 [185] (this article is under the Creative Commons CC BY license; permission is not required as long as the figure is properly cited). A, the two curves of a factor (PVI, peritumoral vascular invasion) are approximately in parallel, indicating the satisfaction of the PH assumption; B, the two curves of a factor (HormRec, hormone receptor status) are not in parallel, indicating the violation of the PH assumption.

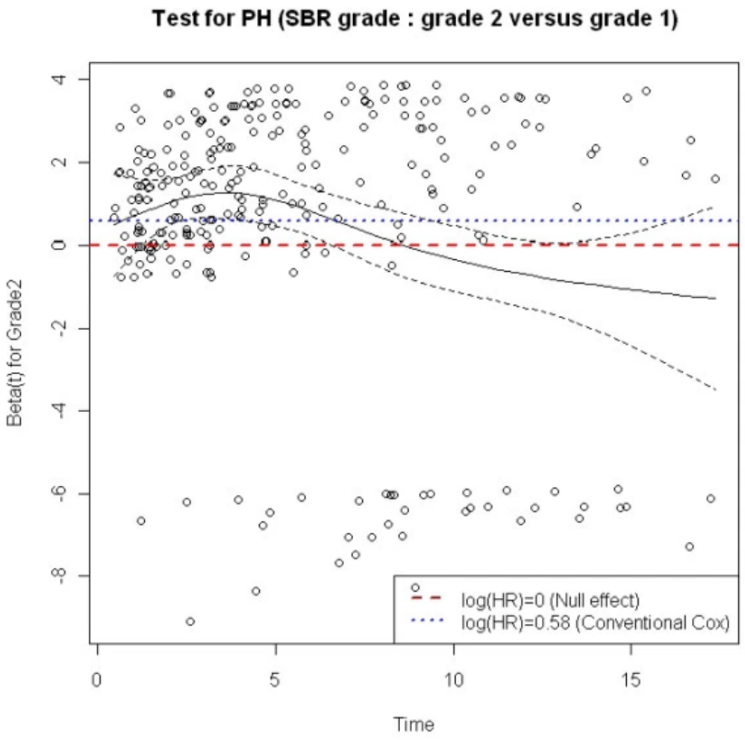


Figure 1.6. A plot of coefficient β(t) based on scaled Schoenfeld residuals for checking the PH assumption. Reprinted from Bellera et al., 2010 [185] (this article is under the Creative Commons CC BY license; permission is not required as long as the figure is properly cited). In this plot, the coefficient (or logarithm of HR) of a factor (Scarff-Bloom-Richardson [SBR] grade) changes over time. In other words, the association of this factor with the outcome of interest varies over time, indicating the violation of the PH assumption. Black dashed lines represent the 95% confidence interval of the coefficient. The dotted line represents the coefficient (or the logarithm of HR) of the factor estimated by the conventional Cox PH model assuming proportional hazards, and the red dash line is the reference line indicating null effect/association.

Unlike graphic methods, the analytic methods provide formal tests for assessing the PH assumption, and they can apply to both categorical and continuous factors. A commonly used method is the score test for trend of slope against time using the scaled Schoenfeld residuals [204]. The cox.zph function in the survival package [205] of R [206] is designed based on this method. Those factors with p values < 0.05, in such a test, are the ones which violate the PH assumption.

Another widely used analytic test is to construct an interaction term between the factor of interest and a function of time f(t), and check whether the coefficient (i.e., logarithm of HR) of the interaction term is equal to zero or not [199,207]. The PH assumption holds when the coefficient is zero, otherwise the assumption is violated for the variable. The function of time can be in different forms, such as the linear, logarithmic, and exponential forms [188,193,199,207,208]. The choice of the time function is important because if it is mis-specified, the factor that violates the PH assumption can be concluded as a factor satisfying the assumption, or vice-versa [199]. Other analytic methods to check the PH assumption can be found in Harrell, 2015 [209] and Ng'andu, 1997 [207].

1.7.2 Modeling time-varying associations

When the PH assumption does not hold for a given factor, the standard Cox PH model is not suitable for examining the association of this factor with the outcome of interest. Instead, other approaches that can accommodate time-varying associations should be used. Such approaches include the stratified Cox PH model, the piecewise/change-point Cox PH model, the Cox PH model with time-varying coefficients, the accelerated failure time (AFT) model, the additive model with time-varying association, the Cox-Aalen model, and the mixture cure model.

1.7.2.1 Stratified Cox PH model

A factor that violates the PH assumption can be fitted into the Cox model by stratifying this factor in the model [210,211]. In this stratified model, different baseline hazard functions are allowed in different strata, and associations of other factors (those that satisfy the PH

assumption) are assumed constant across strata [210,211]. A main limitation of this approach is that it cannot estimate the effect or association of the stratified factor, and thus it is not an appropriate approach if the association of this factor is of the primary interest [211]. Another limitation of this method is that it can only be applied to categorical factors, not continuous ones (although continuous factors can be converted to categorical factors based on certain pre-defined thresholds. In such cases, some information can be partially lost).

1.7.2.2 Piecewise/change-point Cox PH model

In piecewise/change-point Cox PH model, select cut-off time points are used for the factor that violates the PH assumption [210,212]. These cut-off time points, ideally, should be the time points that mark the change (either in direction or magnitude) of association between the factor and the outcome of interest (hence these time points also are referred to as the "change-points") [210,212]. In each time interval defined by these time points, the PH assumption may be satisfied, and thus the association may remain constant (**Figure 1.7**). Between different time intervals, however, associations are different, reflecting the "time-varying" feature of the association (**Figure 1.7**). Other factors that initially satisfy the PH assumption are not assigned with cut-off time points and their associations are checked to make sure they remain constant throughout the follow-up time. The piecewise/change-point Cox PH model provides a straightforward way to deal with time-varying associations [213]. It is also generally regarded as a simple model (especially when there is only one time point assigned) and may also have less assumptions (e.g., do not need to assume a specific baseline hazard function, as do in parametric models [e.g., accelerated failure time model; see Section 1.7.2.4]) compared to other complex models (see the other models below). In addition, compared to the stratified Cox PH model

(which is also a simple model), this approach allows the effect estimation for the factor that violates the PH assumption, and it can be applied to both categorical and continuous factors. For these reasons, the piecewise/change-point Cox PH model was used for accommodating factors of interest that violated the PH assumption in the studies (Chapters 3-5) of this thesis research.

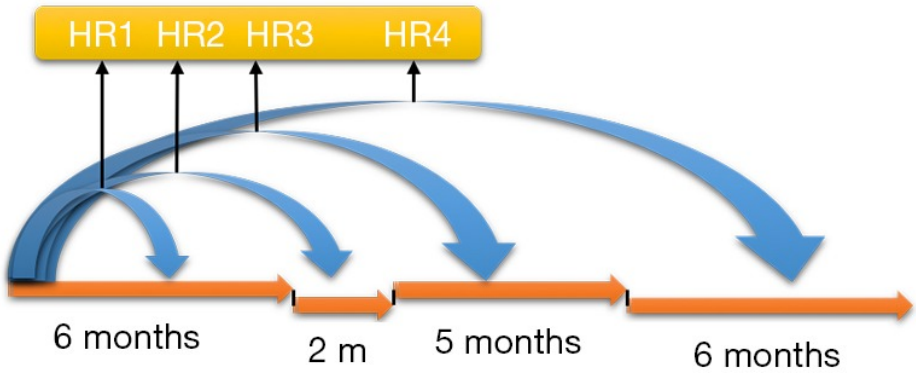


Figure 1.7. Different associations of a factor with survival outcome at different time-intervals post-diagnosis. Reprinted from Z. Zhang et al., 2018 [188] (this article is under the Creative Commons CC BY license; permission is not required as long as the figure is properly cited).

1.7.2.3 Cox PH model with time-varying coefficients

In Cox PH models with time-varying coefficients, coefficient of the factor that violates the PH assumption is a function of time, reflecting the time-varying association of this factor [214]. The piecewise/change-point Cox PH model, theoretically, is a special case of the Cox PH model with time-varying coefficients [188,214], as different associations (or coefficients) in different time intervals can be regarded as the result of a discontinuous time function to the coefficient. In this section, I refer to Cox PH models with time-varying coefficients as those that are not piecewise/change-point Cox PH models.

A common approach to model the time-varying coefficients is to include an interaction term between the factor that violates the PH assumption and a pre-determined function of time [199]. A significant association of the interaction term reveals the time-varying coefficient (also the association) of the factor. Because of that, this model can also be used for checking the PH assumption, as previously mentioned in Section 1.7.1.2. Also as mentioned earlier, the form of the time function is usually unknown, hence choosing the proper form of the function is critical as this may affect the model fitting and study results. A typical form used is the logarithmic function of time [208,210]. Another way to model the time-varying coefficient is to approximate its pattern over time directly from the examined data (using techniques such as spline functions), eliminating the need of choosing a specific time function to the coefficient [193,215,216]. However, for this approach, effectively avoiding over-fitting and under-fitting is a concern [213].

1.7.2.4 Accelerated failure time model

Accelerated failure time (AFT) models are parametric models [217]. Unlike the semiparametric Cox PH model which has its baseline hazard function $h_0(t)$ being unspecified, the AFT models require specified distributions of survival time and thus baseline hazard functions [217]. Based on how the distribution of survival time is specified, the AFT model may take different forms. The most common types of survival distribution in AFT models include exponential, Weibull, log-logistic, and lognormal distributions [217]. Generally, AFT models do not have the PH assumption, and hence AFT models can be used as alternatives to accommodate time-varying associations. The only exceptions are the models based on the exponential and Weibull distributions (the exponential distribution is a special case of the Weibull distribution) in which the PH assumption is automatically satisfied when the AFT assumption is met [218]. The

AFT assumption is an underlying assumption for all AFT models. It assumes that a given factor accelerates or decelerates the survival time of patients by a constant [218,219]. Such a constant is called the acceleration failure factor, which is the measure of association in AFT models (one can think of it as the HR of the Cox PH model).

A critical consideration of using the complicated AFT model for survival analysis is how to choose the proper distribution of survival time (and thus the baseline hazard function) [220]. It is suggested that choosing such a distribution ideally should be based on prior knowledge so that the select distribution is plausible for the problem at hand [221], though assessing the goodness-of-fit of different AFT models can help to determine which distribution is more apporperiate among considered distributions [222].

1.7.2.5 Additive model with time-varying association

While Cox models are multiplicative models in which the combined effects of factors are constructed in a multiplicative form of individual effects, some other models have their combined effects built in an additive way. Such models are called the additive models. Additive models measure the hazard differences (or risk differences), not hazard ratios (or risk ratio), and they do not have the assumption of proportional hazards, making them alternatives to Cox PH models when the PH assumption is violated [223,224]. A commonly used additive model is the Aalen's additive model [225]. In this model, coefficients of all factors are allowed to vary over time, and thus the model can accommodate time-varying associations [225,226]. However, in cases where some factors have time-varying associations while others have constant associations, this model might not be a good choice [223]. Instead, the Cox-Aalen model can be a better option [223].

1.7.2.6 Cox-Aalen model

The Cox-Aalen model is a model that combines the Cox PH and the Aalen's additive models, and hence it is also regarded as a multiplicative-additive model [227]. This model is constructed in a way that the Aalen's additive model replaces the baseline function of the Cox PH model [227]. The Cox-Aalen model is generally considered a more flexible model than the Aalen's additive model, as it accommodates both factors with and without time-varying associations [223]. Of course, this model is also a more complex model than either the Cox PH or the Aalen's additive models.

1.7.2.7 Mixture cure model

The mixture cure model views the study population as a population consisting of long-term disease-free survivors (i.e., those "cured" from the disease; e.g., patients who survived a long time experiencing no disease events) and non-cured survivors (or those who are susceptible to disease events) [228]. For each of the sub-populations, associations of factors are estimated separately but at the same time using different models. For example, the logistic regression model is used to estimate effects of factors for long-term survivors, and the Cox PH or AFT model is used to estimate the associations of factors for short-term survivors [228,229]. Like the Cox-Aalen model, the mixture cure model is also a more complex model compared to the Cox PH model.

1.8 Human genetic variation

Although human genomes share more than 99.9% of their sequences, the remaining < 0.1% of the sequence is different among individuals [230]. Genetic variation refers to such genetic differences. As basic differences among individuals, genetic variations define and/or affect many human characteristics or traits. To better understand their contributions and roles in these characteristics/traits, uncovering the map of genetic variation in the genome was an initial and critical step. Over the past 20 years, different projects were implemented to annotate genetic variations in the human genome, and as of today, hundreds of millions of variants have been identified. The main projects mapping genetic variation include the HapMap, 1000 Genomes, and the gnomAD projects. In general, these projects showed that the majority of the variants in the human genome are non-common variants and are located in non-coding regions (e.g., intergenic regions, introns). A brief summary of these three projects is presented below.

1.8.1 Three main projects mapping genetic variations

The HapMap project was launched by the International HapMap Consortium in 2002 to discover the patterns of human genetic variation [231], empowering researchers to identify variants/genes that affect diseases and other phenotypes. Phase I and II of this project reported around 3 million variants in 270 individuals from four different populations (i.e., Yoruba [YRI]; Japanese [JPT]; Han Chinese [CHB]; and Utah residents with European ancestry [CEU]) [232,233], and the phase 3 (also the final phase) of HapMap expanded the sample size to more than 1,000 (from 11 populations across the world) but only focusing on around 1.6 million variants [234]. The project mainly detected the common genetic variants (MAF ≥ 5%), leaving the majority of rare

variants unidentified [232–234]. Since then, other projects with larger sample sizes have been launched to identify additional genetic variants in the human genome.

The 1000 Genomes project is a landmark, as it created a very comprehensive resource of human genetic variation. This project examined 2,504 individuals from 16 different populations [235], and reported more than 88 million genetic variants in the human genome. Of the 88 million variants, around 75% are rare variants with MAFs < 0.5%, around 14% variants have their MAF between 0.5% and 5%, and about 10% variants with MAF larger than 5% [235]. With regard to the distribution of variants in the genome, the majority (~98%) of them are located in intergenic regions, promoters, introns, and other untranslated regions of genes (e.g., UTR) [235]. Only a small portion of variants are located in gene exons [235].

The Genome Aggregation Database (gnomAD) includes the genetic variation data from a much larger number of individuals (15,708 individuals were subject to whole genome sequencing, and 125,748 individuals were subject to exome sequencing [236]). This project catalogs an unprecedented scale of human genetic variation. More than 200 million variants were identified, including around 15 million variants from the exome dataset [236]. As expected, the number of variants in intergenic and intronic regions is much larger than that of the variants in the exome [236].

Among the variants found in these projects, part of them can have effects/influences on gene expressions and function, and thus potentially on human characteristics/traits. Variants in exons (especially nonsense and missense variants) are expected to be important variants as they may directly lead to aberrant gene expression and function, which can cause or contribute to diseases. The vast majority of non-coding variants may also have regulatory roles on genes and impacts on traits [237,238]. By taking advantage of annotated variants from the described projects

(HapMap, 1000 genomes, and gnomAD), many studies were implemented to identify variants with effects on genes and traits (e.g., diseases), including large scale studies at the genome-wide level. A study that summarised the data from the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/) reported that more than 70,000 variant-trait associations have been identified by GWASs as of September 2018 [239]. Studies on genetic variations mainly focused on SNPs, CNVs, and INDELs, which are the major types of genetic variations in the human genome.

1.8.2 SNPs

A single-nucleotide polymorphism (SNP) is a single-base change in the human genome. Normally, human DNA consists of four basic bases, A, T, C, and G. If a single-base locus in the genome appears with different bases (e.g., A or G) (usually caused by point mutations) among individuals, then this variation is called a SNP.

1.8.2.1 SNPs in the human genome

SNPs are the most common type of genetic variation in the human genome. According to the 1000 Genomes Project, around 83 million out of the 88 million genetic variants identified across populations are SNPs [235]. Consistent with this finding, in the gnomAD data, most (around 89%) variants identified in individuals with the whole-genome sequencing data are the SNPs [236,240]. At the individual genomic level, SNPs are also the most common type of genetic

variation. For a given genome, the total number of variants (including SNPs, CNVs, INDELs, and others) is around 4.1 - 5 million, and more than 85% of them (n = around 3.5 - 4.3 million) are SNPs, according to the 1000 Genomes Project [235]. Because SNPs are so common in the human genome, it is intuitive to assume that they are a major type of genetic source affecting individuals' phenotypes/traits, including diseases, through their effects or influences on gene expression and function.

1.8.2.2 SNPs and gene expression and function

SNPs, both within or outside of the coding regions, can lead to functional consequences, mainly in changing the expression levels and/or normal function of the genes.

SNPs that are located in coding regions, especially nonsense SNPs, can affect normal functions of coded proteins, and thus they are usually considered pathogenic variants and generally have large impacts on phenotypes. Though generally gene functions are affected by SNPs that change the amino acid composition of proteins (e.g., missense variants), this does not mean that silent/sense SNPs (i.e., SNPs in exons that result in synonymous codon substitutions, and hence, no altered amino acids) have no impacts on gene function. For example, it has been reported that three silent SNPs in *MDR1* (encodes P-glycoprotein) can affect the activity of P-glycoprotein, possibly through affecting the translation rate and protein folding and conformation [241]. Other than gene function, SNPs within exons may also affect gene expression levels. An example is the exonic SNPs in autoregulated genes. These SNPs may change amino acids and thus the structures of coded proteins, preventing the binding of these proteins to their DNA binding sites (which usually are located in the promoter regions of coding genes) and the

suppression of further transcriptions [242]. In the end, these proteins are overabundant in cells because of the disruption of autoregulation caused by these SNPs.

While SNPs in exons mostly affect gene function (and maybe gene expression as well), SNPs in non-coding regions (e.g., introns, UTRs, intergenic regions) mainly affect gene expression levels (of course, sometimes they can also affect gene function; e.g., SNPs within alternative splicing sites in introns may result in changed mRNAs and thus non-functional proteins [243]). Such SNPs may regulate/influence gene expression levels at every stage of protein formation, from DNA to post-translation modifications, including chromosomal accessibility, transcription, mRNA splicing, mRNA stability, translation, and protein folding and stability [242]. Generally, SNPs regulate gene expression levels through two main different manners. First, SNPs in enhancers, silencers, insulators, transcriptional factor (TF) binding sites/promoter regions, and other regulatory elements can directly affect the expression of related genes through their impacts on the function of these elements. Second, a SNP may also regulate the expression of a given gene through its effect on another gene [242]. Obvious examples are the SNPs in genes coding miRNAs. These SNPs can affect expression levels of many genes through altering the characteristics or functions of miRNAs [244]. Other than these, considering SNPs may affect epigenetic modifications (e.g., DNA methylation), chromosomal accessibility, mRNA and protein stability, and possible gene-environment interactions, the roles of SNPs on gene expression (as well as gene function) can be highly complex, and more investigations are warranted to fully uncover related mechanisms. Identifying SNPs that associate with mRNA levels (eQTLs) [245,246], DNA methylation (mQTLs; which associate with gene activation) [247–249], DNase hypersensitivity (dsQTLs; which associate with chromosome accessibility) [249,250], TF binding (bQTLs) [249,251,252], and even protein levels (pQTLs) [242] across different tissues at

different developmental stages for different phenotypes/traits are critical steps. Based on these, studies further working on establishing causal links between variants and gene expression/function and detailed regulatory networks can be done. Such studies, therefore, provide important insights into the regulatory architecture of non-coding SNPs in gene expression and function.

As described above, SNPs, both in coding and non-coding region, can influence gene expression or function in different ways. Through such influences, SNPs can have an impact on diseases, including cancers.

1.8.2.3 SNPs and human diseases

As the most abundant genetic variations in the human genome and considering their potential roles in regulating/affecting gene expression and function, SNPs are expected to have impacts on diseases. Studies have shown that many diseases are affected by SNPs, including complex diseases. Unlike the mutations in Mendelian diseases, most SNPs identified in studies on complex diseases are located in non-coding regions and are not in linkage with variants within exons [242,253], reflecting possible regulatory roles of these SNPs on expression and function of disease-causing genes. For example, a common SNP (rs12740374) that is located in 3' UTR of a gene on 1p13 has been known to contribute to the risk of myocardial infarction through its regulatory role on expression levels of *SORT1* (which can then affect the levels of low-density lipoprotein cholesterol [LDL-C], a known risk factor of myocardial infarction) [254]. In colorectal cancer, SNPs in non-coding regions can also affect disease risk. For example, rs6983267, a SNP

within a transcriptional enhancer sequence, has been reported to be able to affect the risk of colorectal cancer, possibly through its regulatory impacts on the oncogene *MYC* [255].

Regarding survival outcomes in cancer specifically, SNPs may also have impacts/influences on these "phenotypes". Though the number of studies (especially large scale GWASs) on survival outcomes in cancer is much less than that of studies on cancer risks, a number of SNPs were found to be associated with survival outcomes in different cancers. For example, a SNP in a miRNA (hsa-mir-196a2) coding sequence was found to be associated with survival of patients with lung cancer [256], and SNPs in genes involved in the telomere pathway were identified to have associations with outcomes of patients with breast cancer [257]. In colorectal cancer, as described in Section 1.6.1.3 and 1.6.2.2, a number of SNPs were also shown to be associated with survival outcomes of patients. Many SNPs identified were only reported for their associations in single survival studies, and among those SNPs that were identified in multiple studies, some showed inconsistent effects (even opposite effect directions). Thus, the associations of these SNPs need further examination. While most of the identified SNPs were reported to have prognostic associations in single types of cancer, some other SNPs were found to be associated with outcomes in several types of cancer. These multi-cancer related SNPs are interesting, as they suggest critical genes and/or pathways contributing to prognoses shared by different cancers. For example, the SNPs in the *VEGF* gene. These variants were found to have associations with survival outcomes of breast, lung, prostate, and colorectal cancers [258,259]. *VEGF* is an important gene in the VEGF pathway, which has a role in angiogenesis and relates to cell proliferation and survival [260]. Such previous studies showed that SNPs can have impacts on cancer outcomes. Some of the SNPs identified as associated with prognosis may not be causal variants. However, these variants may still be informative as they can help pinpoint the causal

variants/genes, especially when the identified SNPs are in high-LD with the causal variants. These identified SNPs, considering their associations with cancer outcomes, can also be potential prognostic markers predicting the outcome risks of cancer patients, and help guide their treatment and follow-up strategies.

As expected, many SNPs that have effects on survival outcomes of cancer patients have related genes/pathways with impacts on or are involved in cell proliferation, invasion, and/or migration. For example, the impact of a common SNP (rs1646724) in the regulatory (promoter) region of *SLC35B4* on tumor recurrence and mortality in prostate cancer is possibly through its effects on *SLC35B4* (encodes glycosyltransferase) expression levels, which influence tumor cell proliferation, migration, and invasion [261]. In colorectal cancer, SNPs in *ERCC5* [262] (encodes an endonuclease with a function in nucleotide excision repair), *MTHFR* [262] (encodes an enzyme participating in folate metabolism), and angiogenesis genes (e.g., *EFNB2*, *MMP2*, and *JAG1*) [175] have also been reported to be associated with disease outcomes, and again, they are involved in cell proliferation, invasion, and/or migration. These pathways make sense as they are able to underscore their roles in tumor growth and/or progression, and hence, patient survival outcomes.

1.8.3 CNVs and INDELs

In addition to SNPs, other major types of genetic variations include CNVs/INDELs. Generally, a DNA segment in a human genome would have two copies, as the human genome is diploid. However, one or both copies of such a segment can be deleted, or it can be replicated into multiple copies (**Figure 1.8**), resulting in a copy number variation (CNV) or

insertion/deletion (INDEL, or indel) variation. While both CNVs and INDELs are DNA segments that vary in copy numbers among individuals, the difference between them is the length of the DNA segment. Though there is no consensus about the segmental length to distinguish CNVs and INDELs, normally, if the length is larger than 1 kb, it is a CNV; otherwise, it is an INDEL. This is also how we define CNVs and INDELs in this thesis research.

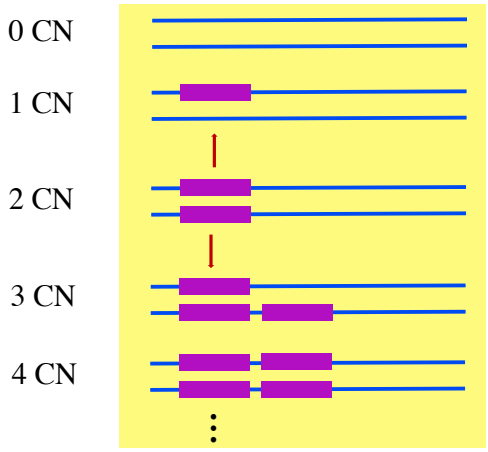


Figure 1.8. CNVs and INDELs are DNA segments varied in copy numbers among individuals. CN, copy number.

1.8.3.1 CNVs/INDELs in the human genome

As mentioned earlier, compared to SNPs, the number of CNVs and INDELs is much smaller. The 1000 Genomes Project captured around 4 million CNVs and INDELs in the human genome [235]. In this data, there are around 600,000 CNVs and INDELs in a given genome [235]. The gnomAD reported a much larger set of CNVs/INDELs, with around 26 million such variants

identified [236,240]. The Database of Genomic Variants (DGV) (http://dgv.tcag.ca/dgv/app/home?ref=) [263] is a curated database of the structural variants with length of > 50bp in the human genome. As of March 18, 2021, there were more than 7 million curated CNVs and INDELs in the database, summarized from 75 studies [263]. The majority of CNVs/INDELs had their length < 50kb, with the most frequent length between 1-10kb [263]. Of course, when short INDELs (i.e., INDELs with length ≤ 50bp) are considered, these variants would account for the largest portion of CNVs/INDELs [235,236,240].

1.8.3.2 CNV/INDEL formation

While SNPs are caused by point mutations, CNVs and INDELs are usually formed under much more complex mechanisms that involve DNA recombination, repair, or replication [264]. These mechanisms can be featured by where the breakpoints of CNVs/INDELs are located in. When their breakpoints are within segmental duplications (SDs), CNVs and INDELs are mostly formed through the mechanism of the non-allelic homologous recombination (NAHR). SDs are DNA segments with lengths > 1kb and high degree (> 90%) of similarity in sequences in the genome [264]. Under the NAHR, two SDs can be aligned to each other, and then crossing over can happen, resulting in deletion or amplification of the affected segment (i.e., CNVs or INDELs) [264]. When the breakpoints are out of the SD regions, this indicates that the corresponding CNVs and INDELs are formed through other mechanisms, mainly the DNA repair- and/or replication-based mechanisms [264,265]. For example, when a template break or replication fork stalling happens during the DNA replication process, it may lead to misconnection/mismatching of strands or switching of the template in the process of DNA repair or replication resumption,

generating CNVs/INDELs [264,265]. As affected segments can include coding or regulatory sequences, generated CNVs/INDELs may then influence gene expression and function.

1.8.3.3 CNVs/INDELs and gene expression and function

Although it is easy to speculate the potential impacts of CNVs and INDELs that include exon sequences (which can lead to aberrant or truncated proteins) or include sequences of entire genes, the knowledge of regulatory functions of CNVs/INDELs that are far away from genes remain to be uncovered. Certain regulatory mechanisms for such variants have been proposed. Similar to SNPs, CNVs/INDELs may also affect gene expression and function at all steps of protein generation/formation.

Generally, CNVs and INDELs in a long distance from genes influence gene expression and function through regulatory elements (e.g., enhancers, repressors, and insulators). Deleting or amplifying a DNA region containing an enhancer or repressor can result in suppression or elevation of mRNA expression levels of the genes [266]. Regarding insulators, CNVs and INDELs can change their presence or numbers, establishing or breaking the regulatory links between other elements and genes. For example, deleting a DNA region containing an insulator which is located between an enhancer and a gene may establish a regulatory link between this enhancer and the gene; on the contrary, an insertion of a DNA sequence containing an insulator between a gene and its enhancer may break their links [267,268]. This way, CNVs and INDELs also lead to differential mRNA expression and protein levels. Similar to SNPs, CNVs/INDELs that contain other non-coding regions (e.g., promoters, introns, transcription factor binding sites) also affect gene expression levels [269,270].

In addition to direct regulatory links between CNVs/INDELs and gene expression/function, CNVs/INDELs may also convey their impacts on genes indirectly (called "trans-regulation"), mainly through diffusible factors. For example, an amplified DNA region with the gene *MYC* (encodes the transcription factor MYC) first elevates protein levels of MYC, which further influences the transcription and translation of MYC-targeted genes [271–273]. Considering some transcription factors (TFs) may be tissue- and/or time-specific (expressed and function in particular tissues at specific times of the development) [274,275], the detailed links between CNVs/INDELs and gene expression and function can be complex. Other than TF-related regulation, earlier mentioned autoregulation is another example of CNVs/INDELs affecting gene expression and function through diffusible factors, though these factors are actually the products of affected genes themselves [242,276]. One of the main differences between direct and indirect regulation is that indirect regulation is generally not "allele-specific" [242]. In other words, in addition to the copy of gene on the same chromosome, a CNV/INDEL with an indirect regulation can also affect the expression and function of the other gene copy on the homologous chromosome in heterozygous individuals, mainly through the affected diffusible factors. Whilst direct regulation of CNVs/INDELs only allows the effects to be imposed on the gene copies on the same chromosomes. Because of that, expression levels of genes may not always be in a linear relationship with the number of effect alleles of CNVs/INDELs (this is also the same as other types of variations, such as SNPs).

CNVs and INDELs also indirectly affect the gene expression and function through their impacts on 3-dimensional (3D) structures and accessibility of genomic regions. An example is that of the CNVs and INDELs covering the boundaries of topologically associating domains (TADs) [277]. TADs are interaction hubs/regions (usually with lengths of megabases) and formed

as DNA loops (boundaries of TADs attach to each other to form loops) in the genome [277]. Within TADs, regulatory elements (e.g., enhancers), spatially, can be brought/pulled closer to their regulated genes and then interact with them [277]. Any changes of the 3D loop structures thus may result in dysregulation of related genes. Therefore, CNVs and INDELs covering TAD boundaries can affect gene expression and protein levels through the changing of the 3D structures of TADs (e.g., by removing or adding TAD boundaries). CNVs/INDELs within a gene involved in chromatin remodeling may also lead to functional consequences. For instance, two INDELs (*BRM*-741 and *BRM*-1321) in the promoter region of *BRM* have been known to be able to affect many genes' expression through their effects on expression levels of *BRM*, which is a subunit of the SWI/SNF complex and involved in chromatin remodeling [278,279]. These two INDELs were examined in Chapter 2.

Thus, CNVs and INDELs may affect gene expression and function in different ways, including direct and indirect ways. Through their impacts on gene expression and function, they naturally, may also have roles in human diseases and traits.

1.8.3.4 CNVs/INDELs and human diseases

Other than the effects of CNVs/INDELs on gene expression and function as mentioned earlier, the potential impacts of these variants on diseases can also be supported by other findings. First, it seems that CNVs/INDELs are more likely to be pathogenic compared to SNPs. According to DGV, 46.13% and 16.18% CNVs/INDELs overlap with transcripts and exons (variations of which are more likely to be pathogenic), respectively [263]. Second, although the number of CNVs/INDELs is much less than that of SNPs, they actually affect more nucleotides

than SNPs in a given genome [235]. Third, CNVs and INDELs seem to have relatively high mutation rates (the mutation rates of CNVs/INDELs can be up to hundreds of thousands of times higher than SNPs), though such mutation rates vary across the genome [265]. High mutation rates may suggest increased chances of causing diseases. While these points indirectly link CNVs/INDELs to diseases, direct relationships between CNVs/INDELs and some diseases have already been shown in research studies.

Currently, CNVs/INDELs are known to be responsible for or associated with a number of Mendelian and complex diseases (for reviews, see Klopocki & Mundlos, 2011 [267], Stankiewicz & Lupski, 2010 [280], and F. Zhang et al., 2009 [265]). Many CNVs/INDELs cause Mendelian diseases, reflecting their large pathogenic effects. For example, the Williams-Beuren syndrome (WBS), which is characterised by vascular stenoses, cognitive deficits, and distinctive facial features, is caused by heterozygous deletion of multiple genes at a locus located in 7q11.23 [281]. Another example is the well-known disease thalassemia. This disease can be caused by deletions of the entire or partial globin genes (e.g., *HBA1, HBA2,* and *HBB*) [265,282–284]. With regard to complex diseases, many CNVs/INDELs contributing disease susceptibilities were identified in studies on neuropsychiatric disorders (such as autism, schizophrenia, and bipolar disease) [285]. For example, deletions and duplications of 1q21.1, 22q11.2, and 15q13.3 contribute to the risks of autism and schizophrenia [265,280]. Other complex diseases affected by CNVs/INDELs include Crohn's disease and different types of cancers. Crohn's disease is a type of inflammatory bowel disease (IBD), and it has been found to be affected by CNVs/INDELs, including a CNV on *HBD-2*/*DEFB4* [286] and a deletion near *IRGM* [287]. In cancers, the two previous mentioned INDELs in the *BRM* gene were found to be associated with the risks of multiple cancers (they were also associated with disease outcomes in different cancers; see Chapter 2), including lung

[95,96], liver [97], and head and neck [96,98] cancers, as well as malignant pleural mesothelioma [99]. All of these showed that CNVs/INDELs have or may play roles in disease pathology.

CNVs/INDELs may also have effects on cancer outcomes. A number of studies have reported that CNVs/INDELs were associated with survival outcomes of cancer patients. For example, amplification in 11q13 has been reported to be associated with a shorter survival time of head and neck cancer patients in several studies (for review, see [288]), possibly because the affected region covers several oncogenes (e.g., *CCND1/bcl-1*, *FGF4/hst-1*, *EMS1/CTTN*). Another example is the *CDKN2A* (a tumor suppressor gene) loss in gliomas. This deletion was found to be associated with worse overall survival of patients with gliomas [289–291]. In colorectal cancer, a number of CNVs/INDELs were reported to be associated with survival outcomes of patients [292,293]. For example, Bi et al. [292] analyzed five CNVs/genes and identified the association of copy number alterations of the *β-TRCP* gene (which encodes a subunit of the Skp1-Cul1-F-box protein complex that functions in ubiquitination) with overall survival in their colorectal cancer patient cohort. Another example is that Lee and others [294] found that the copy number gain of the *MYC* gene was associated with poor prognosis of colorectal cancer patients. Other cancers that have been reported to have associations of CNVs/INDELs with disease outcomes include breast cancer [295], lung cancer [296], bile duct cancer [297], squamous cell carcinoma of the oral tongue [298], esophageal cancer [299], and skull-base chordoma [300]. Note that many CNVs/INDELs with potential influences on outcomes of cancer patients were somatic (or tumor) variations [294,296,298,299,301–304], including the aforementioned 11q13 amplification in head and neck cancer, the *CDKN2A* deletion in gliomas, and the *MYC* copy number gain in colorectal cancer. With regard to germline CNVs/INDELs, an increasing number of studies were performed in

recent years to examine the associations of these variants with outcomes in cancers (mainly breast cancer) [305–307].

Though CNVs/INDELs have been investigated for their relations to diseases, they have not been widely or extensively investigated as SNPs, in both disease risk and outcome research. Part of the reason is that large-scale studies on CNVs/INDELs rely on the map of structural variation, which is much harder to be annotated/described for CNVs/INDELs than SNPs [308]. Many large scale studies on CNVs/INDELs were performed in recent years [285,309–312], by taking advantage of the completion of large projects mapping structural variations (including CNVs/INDELs) in the past several years, especially the 1000 Genomes project [235]. More investigation on CNVs/INDELs is valuable, as it is normally believed that the "missing heritability" explaining unexplained disease-related variance can be further explained by genetic variations, including CNVs/INDELs. Because of that, CNVs/INDELs are attracting more attention in the research community, and a clearer picture regarding their impacts on diseases will hopefully appear in the future. Contributing to this, this thesis research investigated CNVs/INDELs for their associations with disease outcomes (Chapters 2, 3 and 5) in colorectal cancer. In addition, as part of a project in this research, the relations between two *BRM* INDELs and the risk of colorectal cancer were also examined (Chapter 2).

1.9 Rationale and research objectives

Although a number of clinico-demographic factors (e.g., disease stage, tumor location, age at diagnosis) have been established as prognostic markers in colorectal cancer and can predict disease outcomes of patients, there is still a need for additional prognostic markers, as patients with the same disease stage, the same tumor location, and similar age can have varied outcomes. Additional prognostic markers can help stratify patients with high outcome risks from other patients, and such stratification is critical for designing appropriate treatment and follow up strategies for patients.

Many of the previous studies aiming to identify prognostic markers in colorectal cancer focused on clinico-demographic factors, leaving genetic variations (which are also believed to contribute to disease prognosis in colorectal cancer) largely unexamined. Among different genetic variations, SNPs are the most common type but were not substantially investigated, especially on a large scale (e.g., at genome-wide levels). CNVs and INDELs, as other major types of genetic variations that only received more attention in recent years, were even less examined. These genetic variations are thus interesting markers to be investigated for their prognostic value in colorectal cancer.

Additionally, the time-varying associations were rarely investigated for SNPs and CNVs/INDELs, as well as for clinico-demographic factors in colorectal cancer, though factors with such associations are clinically important; they can be early- or late-outcome markers (e.g., tumor recurrence within the first 5 years after diagnosis [early-outcome event]; tumor recurrence after 5 years post-diagnosis [late-outcome event]), respectively. Thus, examining SNPs,

CNVs/INDELs as well as clinico-demographic factors for potential time-varying associations can bring new depth to, and utility in, prognostic research in colorectal cancer.

The main objective of this thesis research was to examine the relationship between genetic/clinico-demographic factors and clinical outcomes in colorectal cancer and to identify factors that associated with patient outcomes. These factors are potential prognostic markers that can predict outcome risks of colorectal cancer patients (including early- and late-outcome markers that can tell the outcome risks in different time periods post-diagnosis). To be more specific, in the study described in Chapter 2, I aimed to test the associations of the two *BRM* INDELs with progression-free survival in colorectal cancer. As part of the study, their associations with disease risk were also examined. In the study described in Chapter 3, the objective was to test the associations between 106 genic CNVs/INDELs and the risk of relapse in colorectal cancer. Later on, in the studies described in Chapters 4 and 5, the updated and long follow up data of the study cohort was used in analyses. The aim of the Chapter 4 study was to examine the associations, including time-varying associations, of clinico-demographic factors with six different outcomes in a colorectal cancer cohort followed up to 19 years. As part of the study, I also described the long-term prognostic characteristics of the cohort. In Chapter 5, the aim was to test the associations of a genome-wide set of SNPs (n = ~4.7 million) and 254 genic and intergenic CNVs/INDELs with disease-specific survival and recurrence/metastasis-free survival in the patient cohort. All these studies were performed based on the data from patients recruited to the Newfoundland Familial Colorectal Cancer Registry (NFCCR).

1.10 Organization of Chapters in the thesis

This thesis is structured in a manuscript style. Other than Chapters 1 (Introduction) and 6 (General Summary and Discussion), Chapters 2-5 are published manuscripts describing my research studies.

This thesis research aims for comprehensive analyses investigating the relationship between genetic/clinico-demographic factors and disease outcomes of colorectal cancer patients, using the data from the NFCCR cohort. Overall, this research significantly contributes to prognostic research in colorectal cancer identifying candidate prognostic makers, including markers with time-varying associations that predict early- or late-outcomes.

My studies, as described in Chapters 2-5, evolved over time based on new knowledge, new data (e.g., updated follow up data), new skills (e.g., imputation), and increased research experience. Chapters 2 and 3 describe studies examining two common types of structural variations (CNVs and INDELs; n = 108) for their associations with disease outcomes in colorectal cancer. The Chapter 2 study investigated two INDELs in the promoter region of *BRM*, a gene that encodes a protein that is involved in chromatin remodeling and thus the regulation of many other genes' expression. These two INDELs are interesting, as they were reported to be associated with outcomes in different cancers, but not in colorectal cancer prior to my study; they were also missed by previous GWASs because they were not included in main genotyping platforms. Thus, investigating these two INDELs in colorectal cancer in the Chapter 2 study was interesting and meaningful. The Chapter 3 study examined a larger number of structural variants (CNVs and INDELs; n = 106) for their prognostic value in colorectal cancer. These CNVs and INDELs were genic CNVs/INDELs, where sequences overlapped with the gene sequences and

were hypothesized to affect gene expression and function (and thus the prognosis of colorectal cancer patients). This study was the first study that investigated a large number of germline CNVs/INDELs for their associations, including time-varying associations, with the risk of relapse in colorectal cancer.

Following Chapter 2 and 3 studies and with the increase of my research experience, in Chapter 5, the study focused on large scale variants including both structural variants (CNVs and INDELs) and non-structural variants (SNPs). By taking advantage of established bioinformatics tools (e.g., IMPUTE2) and based on ~800,000 genotyped SNPs, in the patient cohort, I imputed genotypes of a large number of un-genotyped SNPs. In the end, this study was able to examine a genome-wide set of variants (n = ~4.7 million) for their associations with disease outcomes in colorectal cancer. As part of the study, 254 genic and intergenic CNVs/INDELs (including the 106 genic CNVs/INDELs examined in the Chapter 3 study) were also examined in this study (outcomes of interest were different than that of the Chapter 3 study), by taking advantage of the updated long follow up data of the study cohort (see the next paragraph). This study was the most comprehensive study investigating both structural and non-structural variants in the same patient cohort in colorectal cancer. As time-varying associations were investigated as part of the study, it was also by far the largest scale study examining genetic variants as potential early- or late-outcome markers in colorectal cancer. This study thus uncovered a detailed relationship between genetic variants and disease outcomes in colorectal cancer.

Chapter 4 described a study describing the updated follow up data of the NFCCR cohort. The follow up data of this cohort were updated in 2018 (patients were followed up to 19 years). This long follow up data, which included more outcome events compared to the previous data, made the Chapter 4 and the following Chapter 5 studies have a larger study power compared to

Chapter 2 and 3 studies. Additionally, this long follow up data is preferable in identifying factors with time-varying associations, especially factors that are associated with late-outcomes (i.e., candidate late-outcome markers). As a main part of the study, the Chapter 4 study examined a set of clinico-demographic factors for their associations with six different outcomes of colorectal cancer patients, by taking advantage of the updated long follow up data of the NFCCR cohort. This study thus provided a comprehensive picture of the relationship between clinico-demographic factors and different disease outcomes in colorectal cancer over a long follow up time.

Chapter 6 summarized the results and findings of my research studies that were described in Chapters 2-5 of this thesis. As part of this Chapter, discussions, further directions, and conclusions were also included/presented.

CHAPTER 2: Two functional indel polymorphisms in the promoter region of the Brahma gene (*BRM*) and disease risk and progression-free survival in colorectal cancer

A version of this manuscript has been published in PloS One; 2018, 13(6):e0198873. The manuscript in this Chapter had only minor changes compared to the published version (e.g., "multivariate models" was changed to "multivariable models"). Note that supplementary information that was published with the manuscript is presented in Appendix B.

Yajun Yu[1], Dangxiao Cheng[2], Patrick Parfrey[3], Geoffrey Liu[2,4,5], Sevtap Savas[1,6]

[1] Discipline of Genetics (As of Sep 2020, the Discipline of Genetics has become a part of the Division of Biomedical Sciences), Faculty of Medicine, Memorial University, St. John's, Newfoundland and Labrador, Canada.

[2] Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada.

[3]Clinical Epidemiology Unit, Faculty of Medicine, Memorial University, St. John's, Newfoundland and Labrador, Canada.

[4] Division of Medical Oncology and Hematology, Department of Medicine, Princess Margaret Cancer Centre and University of Toronto, Toronto, Ontario, Canada.

[5] Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada.

[6] Discipline of Oncology, Faculty of Medicine, Memorial University, St. John's, Newfoundland and Labrador, Canada.

2.1 Co-authorship statement

Yajun Yu helped with the study design and methodology, performed all the statistical analyses, interpreted the results, drafted the original manuscript, and reviewed and revised the manuscript.

Dangxiao Cheng performed the DNA genotyping and drafted that part for the manuscript, reviewed and revised the manuscript.

Patrick Parfrey curated the NFCCR clinical demographic, smoking, and survival data, and reviewed and revised the manuscript.

Geoffrey Liu conceptualized the study and reviewed and revised the manuscript.

Sevtap Savas conceptualized, designed, and led the study, and reviewed and revised the manuscript.

2.2 Abstract

Background and objective: The Brahma gene (*BRM*) encodes a catalytic ATPase subunit of the Switch/Sucrose non-fermentable (SWI/SNF) complex, which modulates gene expression and many important cellular processes. Two indel polymorphisms in the promoter region of *BRM* (*BRM-741* and *BRM-1321*) are associated with its reduced expression and the risk of susceptibility or survival outcomes in multiple solid cancers. In this study, we have examined these variants in relation to susceptibility and survival outcomes in colorectal cancer.

Methods: Genotypes were obtained using TaqMan assays in 427 cases and 408 controls. Multivariable logistic and Cox regression models were fitted to examine the associations of the *BRM-741* and *BRM-1321* genotypes adjusting for relevant covariates. Sub-group analyses based on tumor location and sex were also performed. In all analyses, indels were examined individually as well as in combination.

Results: Our results showed that there was no association between the *BRM* polymorphisms and the risk of colorectal cancer. However, genotype combinations of the *BRM-741* and *BRM-1321* variants were associated with the risk of colon cancer. Particularly, patients having at least one variant allele had increased risk of colon cancer when compared to patients with the double wild-type genotype. In the survival analyses, *BRM-741* heterozygosity was associated with longer progression-free survival time in the colorectal cancer patients. A stronger association was detected in the male patients under the recessive genetic model where the homozygosity for the variant allele of *BRM*-741 was associated with shorter progression-free survival time.

Conclusions: Our analyses suggested that *BRM-741* and *BRM-1321* indels are associated with the risk of developing colon cancer and the *BRM-741* indel is associated with the disease progression in colorectal cancer patients, especially in the male patients. Although our results show a different relationship between these indels and colorectal cancer compared to other cancer sites, they also suggest that *BRM* and its promoter variants may have biological roles in susceptibility and survival outcomes in colorectal cancers. Performing further analyses in additional and larger cohorts are needed to confirm our conclusions.

2.3 Introduction

Each year, around 1.4 million people are diagnosed with colorectal cancer and about 700,000 deaths occur because of it [2]. In Canada, around one in 15 people are expected to be diagnosed with this disease in their lifetime [313]. Worldwide, ~40-66% colorectal cancer patients do not survive 5-years after diagnosis [314]. Understanding factors, including genetic factors, that influence the susceptibility to this disease and patient prognosis can help improve its control and patient survival outcomes. For this reason, many studies have examined the associations of genetic variations with the risk of developing colorectal cancer or clinically important events after diagnosis [92,100,105,119,123,176,179,315].

BRM encodes Brahma, one of the two mutually exclusive DNA-dependent ATPase subunits of the SWI/SNF complex [316]. The SWI/SNF complex includes several subunits, exists in multiple forms with different subunit compositions, facilitates transcriptional regulation through remodeling of the chromatin, and is known to play critical roles in many important biological processes, such as cell proliferation and differentiation [317,318]. Not surprisingly, several alterations of the multiple SWI/SNF complex subunits (including of *BRM*) have been identified in cancer, linking them to carcinogenesis or disease progression [279,318].

Loss of BRM is often observed in various types of tumors [98,319–322], which is mainly mediated through epigenetic silencing [319]. Two promoter polymorphisms, *BRM*-741 and *BRM*-1321, are highly correlated with the expression levels of *BRM* [95]. Both of these polymorphisms are indel/repeating sequence variants [95]. *BRM*-741 consists of two (deletion or wild-type allele = Del) or three (insertion or variant allele = Ins) copies of a 7 bp long sequence (TATTTTT) located in 741 bp upstream of the *BRM* transcription start site. *BRM*-1321, on the other hand,

exists as either one (deletion or wild-type allele = Del) or two copies (insertion or variant allele = Ins) of a 6 bp long sequence (TTTTAA) located in further upstream of the *BRM* transcription start site [95]. Variant sequences of these two polymorphisms are highly homologous to the binding site for myocyte enhancer factor-2 (MEF-2), which together with histone deacetylases (HDACs) has been shown to epigenetically silence the *BRM* gene [95,323]. In examination of tissue samples, Liu et al. [95] associated the homozygosity for the variant allele (Ins/Ins) in either or both of the indels with the absence of BRM protein in both lung tumor and unaffected tissues. Examination by Gao et al. [97] showed that in both hepatocellular carcinoma tumors and non-tumor tissues the BRM expression levels decreased similarly with each Ins allele of *BRM*-1321. It is not known at the time being whether the Ins allele of *BRM*-741 has a similar effect on *BRM* expression as in the case of *BRM*-1321 (i.e. expression levels decreasing similarly with each copy of the Ins allele), but considering the fact that the *BRM* silencing is mediated through the binding of the MEF-2 and HDACs to the Ins alleles [95,323], it is a plausible possibility. Last but not least, both indels are linked to each other to varying degrees in different populations (D' = 0.39-0.86) [95,97,98,324–326] and are common in Caucasians with similar minor allele frequencies (MAFs) of 45% [95].

Because *BRM*-741 and *BRM*-1321 can affect the expression of *BRM* and thus the activity of SWI/SNF, it is reasonable to suspect that these two polymorphisms may influence the risk or prognosis of human cancers. Supporting this, specific genotypes of either -741, -1321, or their combinations have been reported to be associated with the risk of lung [95,96], head and neck [96,98], and liver [97] cancers. Similar associations with the survival outcomes of lung [325], esophageal [324], hepatocellular [326] and pancreatic cancer [327] patients have also been detected. However, these two

indels were not evaluated for their potential associations with the risk or survival outcomes in colorectal cancer before. In this study, we tested these associations in colorectal cancer cases and controls from Newfoundland population.

2.4 Methods

2.4.1 Ethical approval

This study was approved by the Health Research Ethics Authority (HREA) of Newfoundland and Labrador (Reference numbers 09.106 and 15.294). Since this was a secondary use of data, no patient consent specific for this study was required.

2.4.2 Study cohorts

Cases and controls recruited to Newfoundland Familial Colorectal Cancer Registry (NFCCR) were examined. NFCCR was described elsewhere in detail [328,329]. In brief, participants (or their family members) provided consent to participate in NFCCR. A total of 750 stage 0-IV cases diagnosed between January 1999 and December 2003 were recruited. Age, sex, and other related demographic information was collected at the time of recruitment. Access to medical records and blood or tissue samples were requested. Individuals free of colorectal cancer were enrolled as controls in the year of 2004 and 2005 by random-digit-dialing [330]. Controls were frequency-matched with the cases in terms of age and sex. In total, 720 controls were recruited.

Blood samples and demographic information using questionnaires were collected at the time of recruitment. Cases who smoked cigarettes before the time of diagnosis were defined as ever-smokers while those did not smoke till this time point were defined as never-smokers. For controls, the time of recruitment was used as the time point to define ever-smoker and never-smoker status. Body mass index (BMI) was calculated based on the body mass and height information provided by the participants. For cases, these data were based on approximately one year before the diagnosis, and for controls, these data were based on approximately two years before their recruitment.

Exclusion criteria for the study cohorts included: (1) cases and controls who were >75 years of age; (2) cases and controls who self-identified themselves as non-white or of mixed race; those who did not provide this information were also excluded; (3) cases who were diagnosed with stage 0 disease; (4) cases who were affected by Lynch syndrome, familial colorectal cancer type X (FCCX), or familial adenomatous polyposis (FAP); (5) cases who were the first, second, or third degree relatives with each other; in such a case one of the patients were randomly excluded. This information was based on a previously obtained genome-wide SNP genotype data [176] and was available for all but three patients; (6) controls who had a known first, second, or third degree relative in the case cohort; (7) controls who are known to have developed colorectal cancer after recruitment; (8) controls with no epidemiological/demographic data; and (9) cases or controls without genomic DNA extracted from blood samples. In the end, 427 cases and 408 controls passed these eligibility criteria. As for the survival analysis, one patient with no prognosis-related data was excluded. Characteristics of the cases and controls are summarized in **Table 2.1**.

Table 2.1. Distribution of baseline characteristics of the study cohorts.

Characteristics	Cases N (%)	Controls N (%)	* P value
Total	427 (100)	408 (100)	
† **Age**			0.40
< 65 years	268 (62.76)	245 (60.05)	
≥ 65 years	158 (37.00)	163 (39.95)	
Unknown	1 (0.23)	-	
Sex			0.91
Female	172 (40.28)	166 (40.69)	
Male	255 (59.72)	242 (59.31)	
Number of FDR with colorectal cancer			**0.0004**
0	304 (71.19)	333 (81.62)	
At least 1	123 (28.81)	75 (18.38)	
Smoking status			0.09
Never	124 (29.04)	143 (35.05)	
Ever	296 (69.32)	265 (64.95)	
Unknown	7 (1.64)	-	
‡ **BMI**			0.35
Underweight and normal	119 (27.87)	127 (31.13)	
Overweight and obese	294 (68.85)	272 (66.67)	
Unknown	14 (3.28)	9 (2.21)	
§ **Disease stage**			-
I	76 (17.84)	-	
II	167 (39.20)	-	

III	140 (32.86)	-	
IV	43 (10.09)	-	
§ **Tumor location**			-
Colon	280 (65.73)	-	
Rectum	146 (34.27)	-	
§ **MSI status**			-
MSS\MSI-L	368 (86.38)	-	
MSI-H	40 (9.39)	-	
Unknown	18 (4.23)	-	
§ **Treatment with adjuvant chemotherapy**			-
No	189 (44.37)	-	
Yes	233 (54.69)	-	
Unknown	4 (0.94)	-	

BMI, body mass index; FDR, first-degree relative(s); MSI, microsatellite instability; MSI-H, microsatellite instability-high; MSI-L, microsatellite instability-low; MSS, microsatellite stable; N, number. P values < 0.05 are bolded.

* Two-sided χ^2 test for comparison between cases and controls with available data.

† Age is the age at diagnosis for cases, and age at recruitment for controls.

‡ Underweight, normal, overweight, and obese are defined as BMI <18.5, 18.5 ≤ BMI < 25, 25 ≤ BMI < 30, BMI ≥ 30, respectively. Categorization criterion was based on the information provided on the website of National Institutes of Health (https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm).

§ Total number of cases is 426.

2.4.3 Follow-up

Patients were followed until the year 2010. The median follow up was 6.98 years (range: 2.00-10.88 years) with 95% confidence intervals (Cis) of 6.69-7.28 years (calculated based on reverse Kaplan-Meier method [331] using IBM SPSS Statistics-23). Data on vital status and occurrence of recurrence and metastasis were collected from various sources as explained in Negandhi et al. [262]. In brief, collection of prognostic data was performed by NFCCR. Clinical events of interest (i.e. recurrence/metastasis/death) were surveyed through and extracted/obtained from the patients' medical records, the Newfoundland Cancer Treatment and Research Foundation database, or patient follow-up questionnaires.

2.4.4 DNA genotyping

All NFCCR cases and controls with available DNA samples were genotyped for the *BRM* promoter indel polymorphisms (n = 493 for cases and n = 448 for controls). Genotyping for *BRM* -741 and *BRM* -1321 promoter region polymorphisms was performed by two custom designed Taqman assays (*BRM*-741: forward primer: 5' TGGCAGGAACGTTCTTTGTG 3'; reverse primer: 5' TGCCGGCTGAAACTTTTTCT 3'; probe for insertion: /56-FAM/TCCCTTTTCTA/ZEN/TTTTTTATTTTTTATTTTTTTTACCTGGAA/3IABkFQ/; probe for wild-type: /5HEX/CCTCCCTTTTC/ZEN/TATTTTTTATTTTTTTACCTGGAAT/3IABkFQ/; *BRM*-1321: forward primer: 5' CATACTTTTCATAACACTACTGCATAGGAACA 3'; reverse primer: 5' TTTTATGAAGTGTGAAAGAATGTTAGGAGACT 3'; probe for insertion: /56-

FAM/A+CT+CTTA+AAAT+T+AAAA+CTGT/3IABkFQ/; probe for wild-type: /5HEX/T+G+CTT+GA+CT+CTTAA+AAC/3IABkFQ/. TaqMan assay reaction condition for *BRM*-741 and -1321 polymorphisms was: 95°C 2 min followed by 40 cycles of 95°C for 6 sec / 60°C for 20 sec. Reaction volume for each sample was 5µl. PCR master mix was obtained from Kapabiosystems (Kapa probe fast qPCR kit, Cat#kk4702). For *BRM*-741 and *BRM*-1321, 9.58% and 19.26% of the DNA samples were genotyped twice and concordance rate was 100%. These methods had been previously compared with Sanger sequencing, and two other sets of primers and probes in 190 patients with 100% concordance.

A total of 831 (n = 424 of 427 cases, n = 407 of 408 controls) and 832 (n = 425 of 427 cases, n = 407 of 408 controls) individuals included into the study were successfully genotyped for the *BRM*-741 and *BRM*-1321 variants, respectively.

2.4.5 Statistical analysis

Hardy-Weinberg equilibrium (HWE) calculations were performed using an online calculator (http://www.oege.org/software/hwe-mr-calc.shtml) [332]. D' and r^2 for linkage disequilibrium (LD) between *BRM*-741 and *BRM*-1321 were calculated by using the LD function of genetics package [333] in R (ver3.2.4) [206]. Chi-squared test was used to examine the differences between cases and controls. All analyses were performed by using R (ver3.2.4) [206] unless otherwise specified.

Similar to other studies, deletions (Del) were assigned as wild-type alleles, and insertions (Ins) as variant alleles. Individual associations of the *BRM*-741 and *BRM*-1321 were analyzed

under different genetic models (co-dominant, dominant, recessive and additive genetic models). Combination analyses involving both polymorphisms were performed as follows: **Category A)** the genotype categorizations used by the previous investigators [95,98,324,325,327]; **Category B)** double homozygous variant genotype (Ins/Ins+Ins/Ins) compared to others; **Category C)** double wild-type genotype (Del/Del+Del/Del) compared to others; and **Category D)** at least one homozygous variant genotype compared to others (shown in **Supplementary Table 1**; Categories A-D).

2.4.5.1 Association analyses

In the case-control study, unconditional logistic regression analyses were applied to test the associations between the indels and the risk of colorectal cancer. Known risk factors (age, sex, and number of first-degree relatives (FDR) with colorectal cancer) were included in multivariable models. Smoking status and BMI were sequentially examined using the log likelihood ratio test. We first examined smoking status and compared models with and without this variable. As the models were significantly different from each other (P values < 0.001) and the model with this variable had a smaller Akaike Information Criterion (AIC) value [334] (and, thus improved model's fit to data), smoking status was included as a covariate in the baseline model. We then examined BMI and obtained similar results. Thus, BMI was also included in the final baseline model as a covariate. Odds ratios (Ors) and 95% Cis for the genotypes were calculated under the multivariable logistic regression models adjusted for the baseline variables.

Cox Proportional Hazards (PH) regression method was used for survival analyses. The outcome of interest was progression-free survival that was defined as the time from diagnosis till the time of death, or, local or distant recurrence. Patients were censored if they experienced none of the events (death, recurrence or metastasis) till the last follow-up. The proportional hazard assumption was tested by using the cox.zph function [205] in R (ver3.2.4) [206]. Age was the only variable that violated the PH assumption (including genotypes), thus multivariable models were stratified by age. Other model covariates included disease stage, tumor location, microsatellite instability (MSI) status, and treatment status (adjuvant chemotherapy Yes/No). Their independent associations with the outcome were confirmed in a multivariable baseline model. Hazard ratios (HRs) and corresponding 95% Cis for the genotype categories were estimated under the age-stratified multivariable Cox models adjusting for these baseline variables.

2.4.5.2 Sub-cohort analyses

To explore whether the associations of these indels vary by sex and tumor location (colon, rectum), we also performed sub-cohort analyses separately (**Supplementary Tables 2-5**). Adjustments in sub-cohort analyses were done by the covariates previously selected, except for the covariate sex in the risk analyses in male and female sub-cohorts, and tumor location in survival analyses in colon and rectal cancer sub-cohorts. In addition, MSI was not included as a covariate in survival analyses of rectal cancer cases because there were only two rectal cancer patients with microsatellite instability-high (MSI-H) tumors.

2.5 Results

2.5.1 Minor allele frequencies, Hardy-Weinberg Equilibrium test, and linkage disequilibrium between the *BRM*-741 and *BRM*-1321 indels

Minor allele frequencies of *BRM*-741 and *BRM*-1321 were 48% and 44% in controls and 47% and 43% in cases, respectively. Both *BRM*-741 and *BRM*-1321 genotype frequencies satisfied the HWE in controls. D' and the r^2 between the *BRM*-741 and *BRM*-1321 were lower than 0.8 in cases (D' = 0.48; $r^2 = 0.20$) and controls (D' = 0.58; $r^2 = 0.29$), indicating the two polymorphisms were not highly correlated with each other in this population.

2.5.2 Associations of the *BRM*-741 and *BRM*-1321 indels with the susceptibility to colorectal cancer

2.5.2.1 Case-control analyses in colorectal cases and controls

Cases (n=427) and controls (n=408) were comparable to each other in terms of frequency distribution of age, sex, smoking status, and BMI, except the number of FDR affected by colorectal cancer as expected (**Table 2.1**). After adjusting for age, sex, number of FDR, smoking status, and BMI, *BRM* indels were not associated with the risk of colorectal cancer when analyzed alone or together (**Table 2.2**).

Table 2.2. *BRM* promoter indels and colorectal cancer risk.

Variable	Genotypes	Cases N (%)	Controls N (%)	OR (95% CI)	* P value
BRM-741					
	Co-dominant model				
	Del/Del (wild-type)	119 (27.87)	113 (27.70)	1 (reference)	
	Ins/Del	215 (50.35)	201 (49.26)	1.09 (0.78, 1.52)	0.61
	Ins/Ins	90 (21.08)	93 (22.79)	0.96 (0.64, 1.44)	0.85
	Unknown	3 (0.70)	1 (0.25)		
	Dominant model				
	Del/Del	119 (27.87)	113 (27.70)	1 (reference)	
	Ins/Ins + Ins/Del	305 (71.43)	294 (72.06)	1.05 (0.77, 1.44)	0.76
	Unknown	3 (0.70)	1 (0.25)		
	Recessive model				
	Ins/Del + Del/Del	334 (78.22)	314 (76.96)	1 (reference)	
	Ins/Ins	90 (21.08)	93 (22.79)	0.91 (0.65, 1.28)	0.59
	Unknown	3 (0.70)	1 (0.25)		
	† **Additive model**				
	Del/Del	119 (27.87)	113 (27.70)	0.99 (0.81, 1.21)	0.90

	Ins/Del	215 (50.35)	201 (49.26)		
	Ins/Ins	90 (21.08)	93 (22.79)		
	Unknown	3 (0.70)	1 (0.25)		
BRM-1321					
	Co-dominant model				
	Del/Del (wild-type)	136 (31.85)	135 (33.09)	1 (reference)	
	Ins/Del	213 (49.88)	188 (46.08)	1.20 (0.87, 1.65)	0.27
	Ins/Ins	76 (17.80)	84 (20.59)	0.93 (0.62, 1.39)	0.73
	Unknown	2 (0.47)	1 (0.25)		
	Dominant model				
	Del/Del	136 (31.85)	135 (33.09)	1 (reference)	
	Ins/Ins + Ins/Del	289 (67.68)	272 (66.67)	1.11 (0.83, 1.51)	0.48
	Unknown	2 (0.47)	1 (0.25)		
	Recessive model				
	Ins/Del + Del/Del	349 (81.73)	323 (79.17)	1 (reference)	
	Ins/Ins	76 (17.80)	84 (20.59)	0.84 (0.58, 1.19)	0.32
	Unknown	2 (0.47)	1 (0.25)		
	[†] **Additive model**				

	Del/Del	136 (31.85)	135 (33.09)		
	Ins/Del	213 (49.88)	188 (46.08)	0.99 (0.81, 1.21)	0.93
	Ins/Ins	76 (17.80)	84 (20.59)		
	Unknown	2 (0.47)	1 (0.25)		
‡ ***BRM-741*** **and** ***BRM-1321*** **genotype combinations**					
	Category A.				
	Double wild-type genotype	73 (17.10)	81 (19.85)	1 (reference)	
	No homozygous variant genotype	223 (52.22)	196 (48.04)	1.36 (0.92, 2.00)	0.12
	One homozygous variant genotype	91 (21.31)	81 (19.85)	1.32 (0.84, 2.08)	0.23
	Double homozygous variant genotype	37 (8.67)	48 (11.76)	0.90 (0.51, 1.56)	0.70
	Unknown	3 (0.70)	2 (0.49)		
	Category B.				
	Other genotype combinations	387 (90.63)	358 (87.75)	1 (reference)	
	Double homozygous variant genotype	37 (8.67)	48 (11.76)	0.71 (0.44, 1.13)	0.15
	Unknown	3 (0.70)	2 (0.49)		
	Category C.				
	Double wild-type genotype	73 (17.10)	81 (19.85)	1 (reference)	
	Other genotype combinations	351 (82.20)	325 (79.66)	1.28 (0.89, 1.84)	0.19

Unknown	3 (0.70)	2 (0.49)		
Category D.				
Other genotype combinations	296 (69.32)	277 (67.89)	1 (reference)	
At least one homozygous variant genotype	128 (29.98)	129 (31.62)	0.93 (0.68, 1.26)	0.63
Unknown	3 (0.70)	2 (0.49)		

CI, confidence interval; Del, deletion; Ins, insertion; N, number; OR, odds ratio.

* Adjusted for age, sex, number of first degree relatives with colorectal cancer, smoking status, and body mass index. Please note that final models include only the patients with the available covariate data. For further information on genotype combinations/categories, please refer to Methods/Supplementary Table 1.

† Ins/Ins vs Ins/Del vs Del/Del.

‡ Homozygous variant genotype is Ins/Ins genotype.

2.5.2.2 Case-control analyses in the sub-cohorts

In multivariable analyses, significant associations were found only in the colon cases and when the *BRM*-741 and *BRM*-1321 indel genotypes were analyzed together (**Table 2.3**). Specifically, compared to double wild-type genotype (Del/Del+Del/Del), no homozygous or one homozygous variant genotypes were associated with the increased risk of colon cancer (no homozygous variant genotype; OR [95% CI] = 1.65 [1.05-2.63]; P value = 0.03; one homozygous variant genotype; OR [95% CI] = 1.77 [1.05-3.01]; P value = 0.03; Category A). Additionally, compared to the double wild-type genotype (Del/Del+Del/Del), combined genotypes that included at least one variant allele were associated with increased risk of colon cancer (OR [95% CI] = 1.60 [1.04-2.50]; P value = 0.03; Category C). There were no associations detected in the rectal cancer, female, or male cancer sub-cohorts (**Supplementary Tables 2** and **3**).

Table 2.3. Associations between *BRM* promoter indels and colon cancer risk.

Sub-cohort	Variable	Genotypes	Cases N (%)	Controls N (%)	OR (95% CI)	* P value
Colon cases + Controls	† ***BRM-741* and *BRM-1321* genotype combination**					
		Category A.				
		Double wild-type genotype	41 (14.64)	81 (19.85)	1 (reference)	
		No homozygous variant genotype	148 (52.86)	196 (48.04)	1.65 (1.05, 2.63)	**0.03**
		One homozygous variant genotype	64 (22.86)	81 (19.85)	1.77 (1.05, 3.01)	**0.03**
		Double homozygous variant genotype	25 (8.93)	48 (11.76)	1.14 (0.60, 2.15)	0.69
		Unknown	2 (0.71)	2 (0.49)		
		Category C.				
		Double wild-type genotype	41 (14.64)	81 (19.85)	1 (reference)	
		Other genotype combinations	237 (84.64)	325 (79.66)	1.60 (1.04, 2.50)	**0.03**
		Unknown	2 (0.71)	2 (0.49)		

CI, confidence interval; N, number; OR, odds ratio. P values < 0.05 are bolded.

* Adjusted for age, sex, number of first degree relatives with colorectal cancer, smoking status and body mass index. Please note that final models include only the patients with the available covariate data. For further information on genotype combinations/categories, please refer to Methods/Supplementary Table 1.

† Homozygous variant genotype is Ins/Ins genotype.

Only the results with P value less than 0.05 are shown in this table; all results obtained in the sub-cohort analyses are shown in Supplementary Tables 2 and 3.

2.5.3 Associations of the *BRM*-741 and *BRM*-1321 indels with progression-free survival in colorectal cancer

2.5.3.1 Survival analyses in the colorectal cancer cases

Results are summarized in **Table 2.4**. The only association was detected under the co-dominant genetic model where the heterozygosity for the *BRM*-741 indel was significantly associated with longer progression-free survival time when compared to wild-type genotype (Ins/Del vs Del/Del; HR [95% CI] = 0.67 [0.45, 0.98]; P value = 0.04; **Figure 2.1**). This association was independent of age, disease stage, tumor location, MSI and adjuvant chemotherapy status.

Table 2.4. *BRM* promoter indels and progression-free survival in colorectal cancer.

Variable	Genotypes	Cases N (%)	P value for PH assumption test	HR (95% CI)	* P value
BRM-741					
	Co-dominant model				
	Del/Del (wild-type)	119 (27.93)		1 (reference)	
	Ins/Del	215 (50.47)	0.72	0.67 (0.45, 0.98)	**0.04**
	Ins/Ins	89 (20.89)	0.95	0.97 (0.62, 1.51)	0.89
	Unknown	3 (0.70)			
	Dominant model				
	Del/Del	119 (27.93)		1 (reference)	
	Ins/Ins + Ins/Del	304 (71.36)	0.86	0.75 (0.53, 1.07)	0.12
	Unknown	3 (0.70)			
	Recessive model				
	Ins/Del + Del/Del	334 (78.40)		1 (reference)	
	Ins/Ins	89 (20.89)	0.81	1.24 (0.85, 1.82)	0.27
	Unknown	3 (0.70)			
	† **Additive model**				
	Del/Del	119 (27.93)			

	Ins/Del	215 (50.47)	0.96	0.96 (0.75, 1.21)	0.72
	Ins/Ins	89 (20.89)			
	Unknown	3 (0.70)			
BRM-1321					
	Co-dominant model				
	Del/Del (wild-type)	136 (31.92)		1 (reference)	
	Ins/Del	212 (49.77)	0.45	0.95 (0.66, 1.36)	0.76
	Ins/Ins	76 (17.84)	0.64	0.98 (0.61, 1.58)	0.93
	Unknown	2 (0.47)			
	Dominant model				
	Del/Del	136 (31.92)		1 (reference)	
	Ins/Ins + Ins/Del	288 (67.61)	0.44	0.95 (0.68, 1.34)	0.79
	Unknown	2 (0.47)			
	Recessive model				
	Ins/Del + Del/Del	348 (81.69)		1 (reference)	
	Ins/Ins	76 (17.84)	0.88	1.01 (0.66, 1.55)	0.96
	Unknown	2 (0.47)			
	[†] **Additive model**				

	Del/Del	136 (31.92)			
	Ins/Del	212 (49.77)	0.54	0.98 (0.78, 1.24)	0.88
	Ins/Ins	76 (17.84)			
	Unknown	2 (0.47)			
‡ ***BRM-741*** **and** ***BRM-1321*** **genotype combinations**					
	Category A.				
	Double wild-type genotype	73 (17.14)		1 (reference)	
	No homozygous variant genotype	223 (52.35)	0.56	0.66 (0.43, 1.00)	0.05
	One homozygous variant genotype	90 (21.13)	0.60	0.72 (0.44, 1.18)	0.19
	Double homozygous variant genotype	37 (8.69)	0.58	1.02 (0.55, 1.89)	0.96
	Unknown	3 (0.70)			
	Category B.				
	Other genotype combinations	386 (90.61)		1 (reference)	
	Double homozygous variant genotype	37 (8.69)	0.60	1.39 (0.81, 2.38)	0.24
	Unknown	3 (0.70)			
	Category C.				
	Double wild-type genotype	73 (17.14)		1 (reference)	
	Other genotype combinations	350 (82.16)	0.79	0.71 (0.48, 1.05)	0.09

Unknown	3 (0.70)			
Category D.				
Other genotype combinations	296 (69.48)		1 (reference)	
At least one homozygous variant genotype	127 (29.81)	0.51	1.07 (0.76, 1.52)	0.69
Unknown	3 (0.70)			

CI, confidence interval; Del, deletion; HR, hazard ratio; Ins, insertion; N, number; PH, proportional hazard. P values < 0.05 are bolded. For further information on genotype combinations/categories, please refer to Methods/Supplementary Table 1.

* Results by age stratified Cox models adjusted for disease stage, tumor location, microsatellite instability (MSI) status, and treatment status (adjuvant chemotherapy Yes or No).

† Ins/Ins vs Ins/Del vs Del/Del.

‡ Homozygous variant genotype is Ins/Ins genotype.

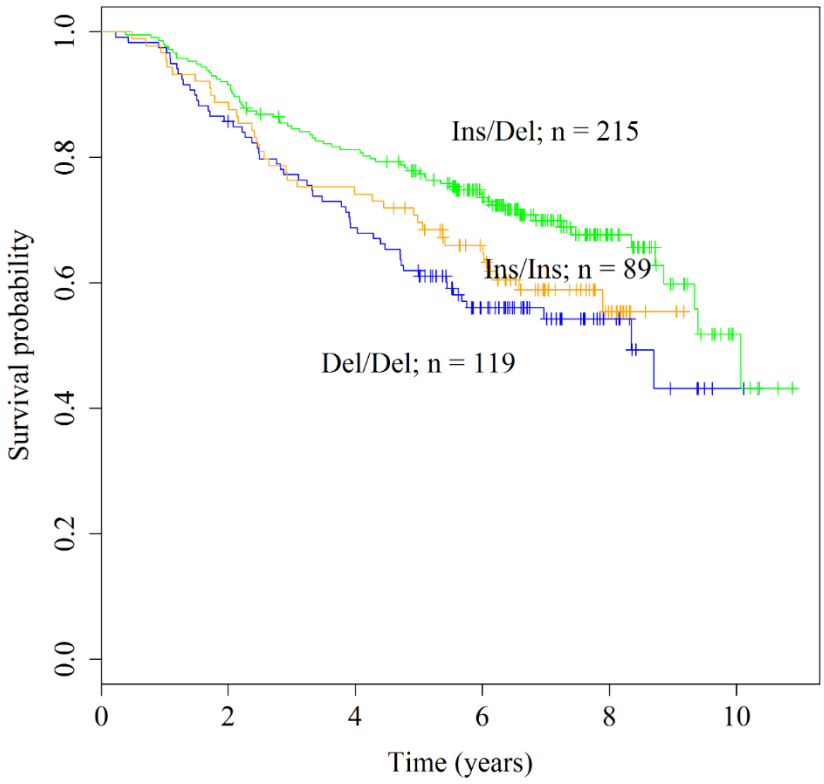


Figure 2.1. Kaplan-Meier curves for the *BRM*-741 indel under the co-dominant genetic model in the colorectal cancer cases. P value of the log-rank test is 0.017.

2.5.3.2 Survival analyses in the sub-cohorts

In the male colorectal cancer cohort, *BRM*-741 indel was associated with the progression-free survival time under the co-dominant and recessive genetic models (**Supplementary Table 4**). Similar to the results obtained in the entire patient cohort (**Table 2.4**), under the co-dominant genetic model heterozygosity for *BRM*-741 was associated with longer progression-free survival time compared to wild-type genotype (HR [95% CI] = 0.54 [0.34, 0.88]; P value = 0.01). This

pattern was also evident in the Kaplan-Meier curves where the male patients with the wild-type Del/Del and the homozygous Ins/Ins genotypes had similar survival probabilities compared to heterozygous Ins/Del individuals who had better survival probability (**Supplementary Figure 1**). A stronger association was detected under the recessive genetic model, where the homozygosity for the *BRM*-741 Ins allele was associated with decreased progression-free survival time compared to other genotypes (HR [95% CI] = 1.84 [1.17, 2.90]; P value = 0.009; **Supplementary Figure 2**). These associations were restricted to the male patients and were not detected in female, colon, or rectal cancer cases (**Supplementary Tables 4** and **5**).

2.6 Discussion

In this study, for the first time we have investigated whether two functional variants (*BRM*-741 and *BRM*-1321) located in the promoter region of the *BRM* gene were associated with the susceptibility to develop colorectal cancer and survival times of the patients. Our results show that presence of at least one variant allele in both of these indels are associated with the increased risk of colon but not rectal cancer. Our results also show that *BRM*-741 may be associated with progression-free survival time in colorectal cancer patients, particularly in the male patients.

BRM codes for one of the two ATPase subunits of the SWI/SNF complex [316]. Two indel polymorphisms in the promoter region of *BRM*, *BRM*-741 and *BRM*-1321, have been shown to be associated with down-regulation of this gene [95], and therefore may affect the function of the

SWI/SNF complex and cellular processes regulated by it. Some of these cellular processes are related to cancer, such as cell proliferation and differentiation [278], making these two genetic variants functionally interesting in cancer research. These variants are not included in many of the genotyping platforms and are not in high LD with other platform polymorphisms to be accurately imputed [95]. This means that the potential associations of these two *BRM* variants may have been missed in genome-wide studies, including one of ours in the NFCCR patient cohort [176]. A number of research groups genotyped and studied the associations of *BRM*-741 and *BRM*-1321 indels with the risk or survival outcomes in various solid cancers. As it is summarized below, while the particular genotypes that are associated with the risk of disease or clinical outcomes are not consistent across different cancer sites, it has been so far consistent that when an association is detected, the variant allele containing genotypes were associated with increased risk of disease/clinical events compared to the homozygous wild-type genotypes. These findings suggest that down-regulation of *BRM* may have a role in carcinogenesis or progression in these cancers.

Studies published so far have showed that either or both of the *BRM*-741 and *BRM*-1321 indels are associated with the risk of development of cancer in multiple, but not all, tissues examined. For example, in stage I-II lung and head and neck cancer patients, one study identified the association of the double homozygous variant genotype with increased disease risk [96]. Two other studies involving stage I-IV patients reported similar associations, in addition to associations of -741 (Ins/Ins genotype) and -1321 (Ins allele containing genotypes), with increased risks of lung cancer [95] as well as head and neck cancers [98]. Additionally, in two separate patient cohorts from Asia, *BRM*-741 was not found to be associated with the disease risk, whereas both the heterozygosity and homozygosity for the *BRM*-1321 Ins allele were

associated with increased risk of liver cancer [97]. However, these associations were not detected in a Canadian cohort in a recent study [326]. Additionally, no associations were found between the two indels and the disease risk in pancreatic cancer (when analyzed either alone or in combination) [327], or in early stage esophageal cancer patients (when analyzed in combination) [96].

In colorectal cancer patients, including the male and female sub-cohorts, our multivariable analyses detected no associations of the *BRM*-741 and *BRM*-1321 indels with the disease risk when these variants were analyzed individually or in combination. However, when the analysis was restricted to the colon cancer patients, genotypes containing at least one variant allele were associated with increased risk of colon cancer compared to the double wild-type genotype (Del/Del+Del/Del) (**Table 2.3** – Category C). These associations were independent of age, sex, number of first degree relatives with colorectal cancer, smoking status, and body-mass index. Additionally, as also shown in **Table 2.3** (Category A), no homozygous and one homozygous genotypes were associated with increased risk of colon cancer compared to double wild type genotype. While we have not observed the association of the double homozygous variant genotype with increased cancer risk compared to double homozygous wild type genotype (likely because of the rarity of this genotype in our cases and controls; **Table 2.3** – Category A), the fact that the presence of the variant alleles associates with increased colon cancer risk is biologically in line with the findings in other cancers. Therefore, similar to other cancer sites (e.g. lung and head and neck cancers [95,96,98]) our results suggest that the loss or reduced expression of *BRM* may increase the colon cancer risk. Interestingly, another gene coding for a subunit of the SWI/SNF complex (*ARID1A*/*BAF250A*) has been found to have frameshift or nonsense mutations in up to 10% of colon tumors [335], suggesting that abnormalities in ARID1A protein may have a role colon carcinogenesis. Together with our results, these findings suggest

the possible involvement of the SWI/SNF complex in colon carcinogenesis. Overall, once confirmed in other patient cohorts our results may have significant implications for understanding the biological functions of the *BRM* gene and the SWI/SNF complex, and their potential roles in pathogenesis or treatment of colon cancer. In contrast, there was no evidence of associations of the *BRM* indels with the risk of rectal cancer. This may be attributed to insufficient power in the rectal cancer cohort (n=146), or the fact that colon and rectal cancers are separate cancer sites arising in distinct tissues characterized with different pathogenesis and molecular alterations [336]. Further cohort and/or molecular studies can be valuable in addressing this hypothesis.

Similar to susceptibility studies, associations of the *BRM*-741 and *BRM*-1321 indels with survival outcomes have been reported in multiple cancer sites. For example, in pancreatic as well as in esophageal cancers, one or two copies of the indel variant alleles (Ins/Del, Ins/Ins) or double homozygous variant genotype (Ins/Ins+Ins/Ins) were associated with reduced overall survival time [324,327]. Additionally, in two separate stage III-IV non-small cell lung cancer cohorts, homozygosity for the variant alleles of either indels as well as the double homozygous variant genotype were associated with shortened overall and progression-free survival time [325]. A recent study on liver cancer patients also showed similar associations between overall survival time and these indel variants [326]. In our study, no associations were detected between the progression-free survival time of the patients and the *BRM*-1321 genotypes or the genotype combinations of the *BRM*-741 and *BRM*-1321 indels. However, associations were detected for the *BRM*-741 genotypes. Specifically, when compared to the wild-type genotype (Del/Del), heterozygosity for the *BRM*-741 indel was associated with longer progression-free survival time in the colorectal cancer cohort independent of age, disease stage, tumor location, MSI and adjuvant treatment

status (**Table 2.4**; **Figure 2.1**). A similar association was also detected in the male colorectal cancer patients (**Supplementary Table 4**; **Supplementary Figure 1**). Based on the previous studies on other cancers, we would expect the wild-type genotype to have better survival outcomes compared to the genotypes that include the variant allele. However, our results do not support this assumption. We also note that it is possible that the small sample size in the wild-type homozygous genotype group may have led to missing a possible association. In addition, under the recessive genetic model we found that the male patients who had the Ins/Ins genotype of *BRM*-741 had shorter survival times compared to the rest of the male patients (**Supplementary Table 4**; **Supplementary Figure 2**). This association was not detected in the female patients ($p > 0.05$: **Supplementary Table 4**). However, as shown in **Supplementary Figure 3**, while it did not reach significance, an opposite effect of the Ins/Ins allele in the female patients was observable, suggesting that the prognostic associations of the *BRM*-741 may be different between male and female colorectal cancer patients. This opens new research avenues for future studies that can help dissect the biological basis of sex-based differences in colorectal cancer outcomes.

Strengths/limitations of this study can be summarized as follows: replications in independent patient cohorts are required to rule out false-positive associations and to confirm our results; death from any cause was used as one of the endpoints as the cause of death information was not available for all patients; and the low frequency of the double homozygous variant genotype has possibly prevented examination/detection of its potential associations in our cohort, thus analysis of larger patient cohorts are needed. However, to our knowledge, this is the first study that investigated the association between the *BRM*-741 and *BRM*-1321 promoter variations and disease risk and patient survival outcomes in colorectal cancer; the patient cohort is a well

described cohort with long follow-up time (median: 6.98 years); a comprehensive investigation has been conducted including application of multiple genetic models and sub-group analyses; and more importantly, in the survival analysis the proportional hazard assumption of the Cox regression method has been assessed and appropriate models have been constructed, which makes our estimations more reliable [184,185].

2.7 Conclusions

In conclusion, our results suggest the potential involvement of *BRM* in colon cancer pathogenesis and colorectal cancer progression. Analyses in larger and additional patient cohorts are needed to verify our results.

2.8 Acknowledgements

We are grateful for the patients and controls that participated in the NFCCR and made this study possible; Dr. Roger Green who was involved in the initial phases of this study, who passed away before the end of study; Andrea Kavanagh for extracting the epidemiological data from the NFCCR database; Drs. Elizabeth Dicks and Jane Green and many other investigators and staff for their efforts in collection of the participants and their information; and Megan Carey for helping with the tables. SS is a Beatrice Hunter Cancer Research Institute (BHCRI) senior investigator.

CHAPTER 3: Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying associations with the risk of relapse

A version of this manuscript has been published in Cancer Medicine; 2017, 6(6):1220-1232. The manuscript in this Chapter had only minor changes compared to the published version (e.g., "time-varying effects" was changed to "time-varying associations". This is for keeping consistency of terminology throughout the thesis). Note that supplementary information that was published with the manuscript is presented in Appendix C.

Salem Werdyani[1], Yajun Yu[1], Georgia Skardasi[1], Jingxiong Xu[2], Konstantin Shestopaloff[3], Wei Xu[2,3], Elizabeth Dicks[4], Jane Green[1,5], Patrick Parfrey[4], Yildiz E. Yilmaz[1,4,6] & Sevtap Savas[1,5]

[1] Discipline of Genetics (As of Sep 2020, the Discipline of Genetics has become a part of the Division of Biomedical Sciences), Faculty of Medicine, Memorial University, St. John's, Newfoundland and Labrador, Canada

[2] Department of Biostatistics, Princess Margaret Hospital, University of Toronto, Toronto, Ontario, Canada

[3] Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, Canada

[4] Clinical Epidemiology Unit, Faculty of Medicine, Memorial University, St. John's, Newfoundland and Labrador, Canada

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

[6] Department of Mathematics and Statistics, Faculty of Science, Memorial University, St. John's, Newfoundland and Labrador, Canada

Authors, including the first author Salem Werdyani, as well as the journal Cancer Medicine have given permissions to include this manuscript in the current thesis. The work or contributions that I (as well as other authors) have done or made were clearly stated in the following co-authorship statement. To be more clear, sections that include my work or contributions were also marked and explained as footnotes in this chapter. The School of Graduate Studies of Memorial University of Newfoundland also agreed with including this manuscript (in which I [Yajun Yu] am the second author) in this thesis as long as the authors and the journal gave permissions, and the work and contributions of authors were clearly stated (which I have done).

3.1 Co-authorship statement

Yajun Yu performed the survival analysis examining the associations between CNVs/INDELs with relapse-free survival in colorectal cancer, interpreted the results of survival analysis, helped draft the manuscript.

Salem Werdyani performed the CNV and INDEL predictions and pathway analysis, identified the variants unique to FCCX cases; and helped draft the manuscript.

Georgia Skardasi performed the duplex PCR analyses; helped with literature search and statistical analyses; helped draft the manuscript.

Jingxiong Xu, **Konstantin Shestopaloff** and **Wei Xu** performed the initial quality control and population stratification analyses on the patient cohort.

Elizabeth Dicks helped collect the patients and their clinical and outcome data.

Jane Green helped collect the patients/families at NFCCR.

Patrick Parfrey established and led the NFCCR; led collection of patients, and their clinical and outcome data.

Yildiz E. Yilmaz supervised the statistical analyses.

Sevtap Savas conceived and led the project; supervised all analyses; helped draft the manuscript, and finalized and submitted the manuscript.

3.2 Abstract

INDELs and CNVs are structural variations that may play roles in cancer susceptibility and patient outcomes. Our objectives were a) to computationally detect and examine the genome-wide INDEL/CNV profiles in a cohort of colorectal cancer patients, and b) to examine the associations of frequent INDELs/CNVs with relapse-free survival time. We also identified unique variants in 13 Familial Colorectal Cancer Type X (FCCX) cases. The study cohort consisted of 495 colorectal cancer patients. QuantiSNP and PennCNV algorithms were utilized to predict the INDELs/CNVs using genome-wide signal intensity data. Duplex PCR was used to validate predictions for 10 variants. Multivariable Cox regression models were used to test the associations of 106 common variants with relapse-free survival time. Score test and the multivariable Cox proportional hazards models with time-varying coefficients were applied to identify the variants with time varying effects on the relapse-free survival time. A total of 3,486 distinct INDELs/CNVs were identified in the patient cohort. The majority of these variants were rare (83%) and deletion variants (81%). The results of the computational predictions and duplex PCR results were highly concordant (93-100%). We identified four promising variants significantly associated with relapse-free survival time (p-values < 0.05) in the multivariable Cox proportional hazards regression models after adjustment for clinical factors. More importantly, two additional variants were identified to have time-varying associations with the risk of relapse. Finally, 58 rare variants were identified unique to the FCCX cases; none of them were detected in more than one patient. This is one of the first genome-wide analyses that identified the germline INDEL/CNV profiles in colorectal cancer patients. Our analyses identified novel variants and genes that can biologically affect the risk of relapse in colorectal cancer patients. Additionally, for the first time we identified germline variants that can potentially be early-relapse markers in colorectal cancer.

3.3 Introduction

Colorectal cancer is the third most commonly diagnosed cancer and the fourth leading cause of cancer related deaths worldwide [2]. Both the incidence and mortality rates of this disease show variability around the world; the incidence rates are higher in developed countries, such as Japan, Australia/New Zealand, USA, Europe, and Canada [5,6]. Despite a higher rate of incidence, interestingly, the survival rates are generally much better in the developed countries compared to developing countries. For example, the 5-year survival rate of colorectal cancer patients is around 65% in the USA and Canada, which is higher than the survival rates in developing countries [5,337]. The root cause of this geographic disparity is unknown, but variable lifestyle, socioeconomic, or environmental factors, or widespread screening and diagnostic programs in developed countries compared to the developing countries are suspected factors [5,6]. In addition to these factors, genetic factors may also influence the risk of susceptibility and disease outcomes in patients. The promise of the *personalized medicine* is that such genetic factors influencing the susceptibility may be used for prevention and screening purposes, while those predicting the prognosis may be used to predict the potential course of the disease, and thus, to inform the treatment decisions [338,339].

Among the genetic factors are the structural variants, such as insertion/deletion (INDEL) and copy number variation (CNV) polymorphisms [340,341]. Both INDELs and CNVs are DNA segments that present at variable copy numbers (i.e. caused by deletions or insertions/amplifications) among the individuals of a population. Both types of variants can also be inherited or formed *de novo*. Yet the main difference between the INDELs and CNVs is their sizes: while there is no consensus, usually those variants shorter than 1 kb are called INDELs,

whereas larger variants are called CNVs. Compared to single-nucleotide polymorphisms (SNPs), the most common type of genetic variation in the human genome, structural variations (with the exception of 1 bp INDELs) affect more nucleotides [340] and are characterized by a higher per-locus mutation rate, and thus these variants are considered to be a major source of genetic as well as phenotypic variability in humans [341,342]. A significant portion of INDEL/CNV sequences also contain parts or the entire sequences of genes (i.e. genic INDELs/CNVs), and hence may affect gene function or expression [340,341]. Understandably, such biological effects may lead to alteration of human physiological functions, which may contribute to the pathogenesis or progression of human diseases. In fact, an increasing number of studies have shown the associations or roles of INDELs/CNVs in both Mendelian and complex diseases, including cancer [343–345].

In colorectal cancer, a small number of studies examined the germ-line (i.e. non-tumor DNA) INDELs/CNVs and their links to disease susceptibility, including hereditary colon cancer syndromes such as Familial Colorectal Cancer Type X (FCCX) [346–349]. A number of studies also looked at the associations of deletion of select genes (such as *GSTM1*, *GSTT1*) with the disease outcome [262,350,351]. However, a comprehensive identification of INDELs/CNVs in a large patient cohort and their examination in relation to survival outcomes have not been done before. In this study we aimed to detect the germline INDEL/CNV profiles in a colorectal cancer patient cohort and to test the possible associations of common and genic INDELs/CNVs with the patient relapse-free survival times. We also identified the rare INDELs/CNVs that are only detected in patients diagnosed with FCCX.

3.4 Materials and Methods

3.4.1 Ethics approval

This study was approved by the Health Research Ethics Authority (HREA) of Newfoundland and Labrador (Reference numbers 09.106, 13.073 and 15.294).

3.4.2 Patient cohort and the genome-wide data

The patient cohort examined in this study was previously described [176]. In short, it included 505 patients out of 750, who were recruited to the Newfoundland Colorectal Cancer Registry (NFCCR) between January 1999 and December 2003 [328,329]. A written consent and permission to access tissues and medical reports were obtained from patients or their close relatives. Peripheral blood samples were collected from most of the patients at the time of recruitment and were used to extract genomic DNA. Patient follow up was performed as described by Negandhi and his co-authors [262]. Among 750, 539 stage I-IV patients with available clinicopathological and outcome data as well as germ-line (i.e. blood-extracted) DNA samples were genotyped (service provider: Centrillion® Biosciences, CA, USA) using the Illumina® Human Omni1_Quad_v1 genome-wide SNP genotyping platform, as reported previously [176]. This high-resolution Illumina Infinium® BeadChip is designed to provide the genome-wide SNP genotype, as well as the signal intensity data for 1,140,419 probes (http://www.illumina.com/documents/products/datasheets/datasheet_humanomni1_quad.pdf). In this study, the signal intensity data for each patient was used as input for detection of their

INDELs/CNVs. Probe locations in this platform were based on the human genome coordinate 19 (hg19), which was used throughout this project.

Subsequent to the SNP genotyping reaction of 539 patients, a set of quality control and population structure analyses was carried out as reported earlier [176]. At the end, 505 Caucasian and unrelated patients constituted the initial, starting cohort in this study.

3.4.3 Detection of INDELs/CNVs

The main steps used to detect INDELs/CNVs in this study are summarized in **Figure 3.1**. Variants were detected using two different algorithms, QuantiSNP [352] and PennCNV [353], followed by a series of quality control and exclusion criteria as described in detail in **Appendix C**. A total of 495 patients out of the initial set of 505 patients had satisfied these criteria, and thus, formed the final study cohort (**Table 3.1**).

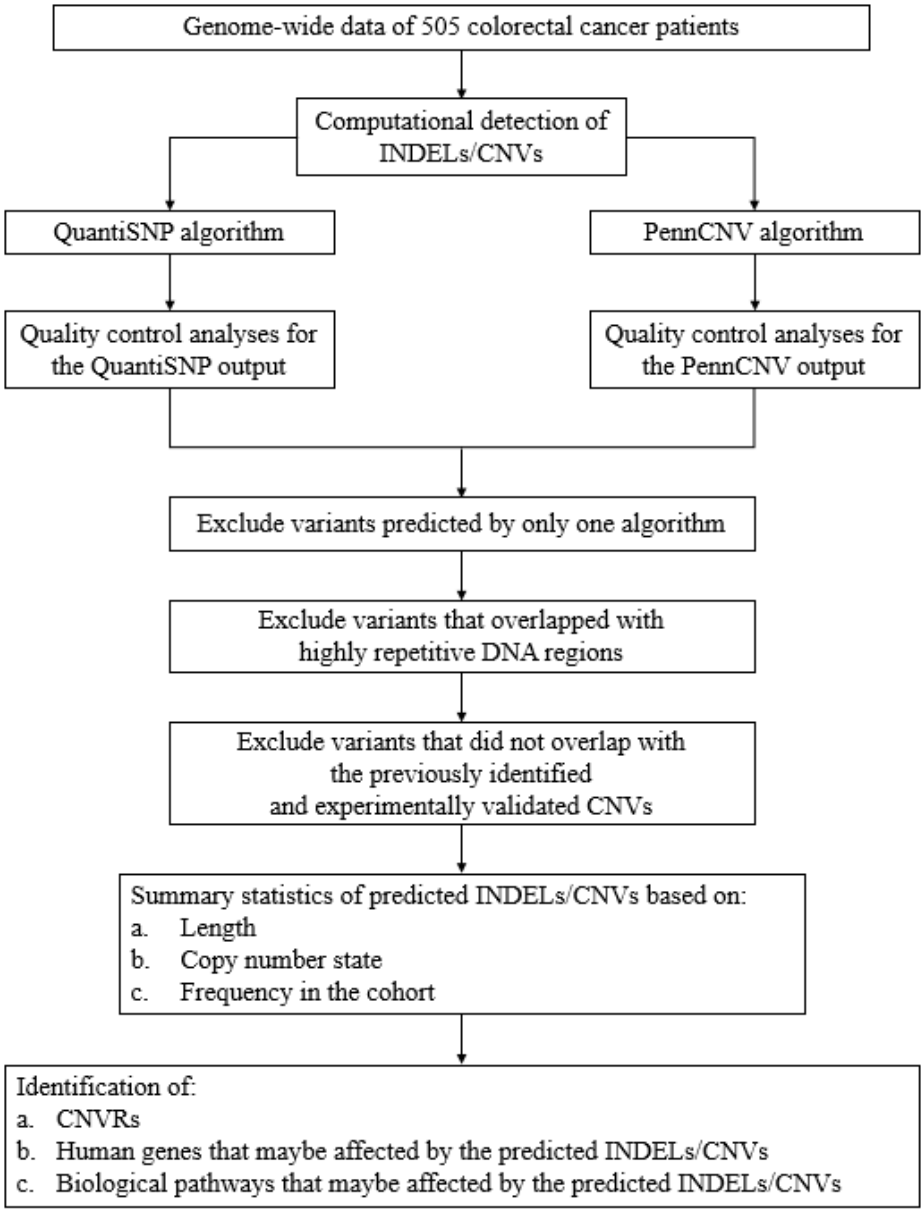


Figure 3.1. The main steps of the computational analysis that were used to detect, describe, and examine the INDELs/CNVs in the patient cohort. CNV: copy number variation; CNVR: CNV region; INDEL: insertion/deletion.

Table 3.1. The baseline features of the patient cohort.

Features	Number	%
Sex		
Female	194	39.19
Male	301	60.81
Age at diagnosis		
<65	312	63.03
≥65	183	36.97
Location		
Colon	328	66.26
Rectum	167	33.74
Histology		
Non-mucinous	438	88.48
Mucinous	57	11.52
Stage		
I	89	17.98
II	193	38.99
III	164	33.13
IV	49	9.90
Grade		
Well/moderately differentiated	457	92.32

Poorly differentiated	34	6.87
Unknown	4	0.81
Vascular invasion		
Absent	300	60.61
Present	158	31.92
Unknown	37	7.47
Lymphatic invasion		
Absent	290	58.59
Present	166	33.54
Unknown	39	7.88
MSI status		
MSI-L/MSS	421	85.05
MSI-H	53	10.71
Unknown	21	4.24
Tumour *BRAF* Val600Glu mutation		
Absent	402	81.21
Present	47	9.49
Unknown	46	9.29

MSI-H: microsatellite instability-high; MSI-L: microsatellite instability-low, and MSS: microsatellite stable.

3.4.4 Identification of genes and biological pathways possibly affected by the INDELs/CNVs

To identify the genes that are possibly affected by the INDELs/CNVs, an overlap (≥1 bp) analysis was performed between the distinct INDELs/CNVs and the list of expressed sequences based on the hg19 that was obtained from the ENSEMBL database on August 2014 [354]. These INDELs and CNVs are called as "genic INDELs and CNVs" throughout this study. In order to obtain the protein pathway information the list of genes that overlapped with the INDELs/CNVs was loaded into the "Gene List Analysis" tool of the PANTHER database [355] on September 2015.

3.4.5 Experimental validation of select INDELs/CNVs

Selection of CNVs: For DNA analysis, we prioritized those INDELs/CNVs that were homozygously deleted in at least 5% of the patients. Whenever possible, we aimed to further prioritize INDELs/CNVs that overlap/delete the sequence of an entire gene over those that partially overlap with genes. A literature search was also performed and functional relevance to cancer was also considered. At the end, 10 INDELs/CNVs that affect the sequences of *ADAM3A/ADAM5A*, *CNOT1*, *DLEU1*, *FAM149A*, *FILIP1L/CMSS1*, *LCE3C/LCE3B*, *NME7*, *REV1*, *WDR34/VTI1BP4*, and *WWOX* genes were selected for experimental validation.

Duplex end-point PCR: Duplex end-point PCR was performed for selected genic INDELs/CNVs in the DNA samples of 100 of the patients. This analysis can distinguish between the patients with homozygous deletion and those with at least one copy of the variant. We opted

for duplex PCR rather than quantitative methods due to availability of low amount of patient DNA samples. Oligonucleotides and amplification conditions are described in **Appendix C**.

3.5 Statistical analyses [i]

All statistical analyses were performed by R (version 3.2.4) [206] or SPSS (IBM-SPSS versions 22 and 23).

A) INDELs/CNVs: The 106 variants (31 INDELs and 75 CNVs) with the following features were selected for survival analyses: i) INDELs/CNVs whose sequences overlap with genes (i.e. genic INDELs/CNVs), and ii) INDELs/CNVs that had at least 10% (while also not exceeding 90%) of the patients with the copy number state (CN) of 0. Our hypothesis was that patients who were homozygously deleted for the CNV/INDEL sequence (and thus likely have both copies of the gene affected; CN = 0) had different survival outcomes than those patients had at least one copy of the INDELs/CNVs (and thus with at least one copy of the gene unaffected by the INDELs/CNVs; CN ≥ 1). Hence, during the statistical analyses, patients were categorized as CN = 0 versus CN ≥ 1, where the latter group of patients served as the reference group. Information related to these CNVs/INDELs and genes are shown in **Supplementary Table 6**.

B) Survival outcome: Relapse-free survival (RFS) was defined as the time from diagnosis till the time of diagnosis of local or distant recurrence (i.e. metastasis), or death (whichever occurred earlier). Patients who did not experience these events were censored at the time of their last

[i] The work described in this section was done by the thesis author (i.e., Yajun Yu).

contact. For two out of 495 patients, either the relapse status or the relapse/last contact date was missing. During the entire follow up period, a total of 197/493=40% of the patients have experienced relapse.

C) Baseline variables and survival analyses: Potential multicollinearity among the baseline variables was checked using the Pearson's correlation test in R. As a result, vascular and lymphatic invasion were found to be highly correlated with each other (r^2=0.96); between the two, the one with the smaller number of missing values (i.e. vascular invasion) was included into the baseline modeling.

Survival analyses were done using the survival package in R [356]. We first tested the associations of variables with RFS assuming all variables satisfied the proportional hazards (PH) assumption of the Cox PH regression model. We also tested the PH assumption for each variable and, when appropriate, modeled survival outcome using the Cox regression model with time varying coefficients.

i) Survival analysis assuming all variables satisfied the proportional hazards (PH) assumption of the Cox PH regression model

Univariable Cox PH regression model was fitted for each baseline variable; those that had a p-value less than 0.1 were then analyzed in a multivariable Cox PH regression model (stage, location, sex, vascular invasion, and microsatellite instability [MSI]). Variables that remained significant in this model were disease stage, tumor location, and MSI status. We confirmed the independent associations of these variables (stage, MSI, and tumor location) with RFS in a separate model that only contained these variables. Genotypes of each INDEL/CNV were then adjusted for these baseline variables in Cox PH regression models using the coxph function in R (**Supplementary Table 7**).

ii) Testing the PH assumption for each variable and, when appropriate, modeling survival outcome using the Cox regression model with time varying coefficients.

We used the score test [204] to check whether the study variables violated the PH assumption (i.e. the hazard ratio does not remain constant suggesting that the effect of the variable on the RFS changes over time). Among the baseline variables in **Table 3.1**, age at diagnosis (defined as < 65 years of age versus ≥ 65 years of age) was the only one that violated this assumption. Thus, we first examined the baseline variables that had a p-value < 0.1 in the univariable analyses (stage, sex, vascular invasion, location, and MSI) in an age-stratified Cox PH regression model. As a result, disease stage, tumor location, and MSI status remained significant. Thus, the final baseline model consisted of age as stratum and disease stage, MSI status, and tumor location as variables for adjustment. Associations of each of the 106 INDELs/CNVs with RFS were then examined in these models with or without time varying coefficients as appropriate. To do so, we first examined each of the variants using the score test [204] under the stratified multivariable models to evaluate whether any of them violated or satisfied the PH assumption. Variants that satisfied the PH assumption were investigated in age stratified conventional Cox PH regression models (without the time-varying coefficients) (**Supplementary Table 8**). For those variants that violated the PH assumption (i.e. potential variants with time-varying associations; score test $p < 0.05$), we first estimated the time-point before and after which their effects on the RFS changed by following the approach described by Pavelitz and others [196]. In brief, we considered each of the time-points (and used the survSplit and cox.zph functions in R) starting with $t_1 = 0.1$ with 0.1 year increments till the end of follow-up time (10.8 years) in age-stratified multivariable models. The time-point at which the model had the largest maximized log partial likelihood was deemed to be the time-point where the

effect of the variants on RFS changed [196]. Score test was again applied to check the PH assumption before and after the identified time-point for each variant and the coxph function was used to estimate the hazard ratios and confidence intervals for these time periods.

A p-value < 0.05 was assumed significant. Because of the exploratory nature of this study and in order to limit false-negative results, a correction for multiple testing was not performed.

3.6 Results

3.6.1 Characteristics of the distinct INDELs/CNVs

Baseline characteristics of 495 patients whose data passed the quality control thresholds by both QuantiSNP and PennCNV algorithms and who constituted the final cohort of patients are summarized in **Table 3.1**.

Collectively, in all patients 3,486 distinct INDELs/CNVs (**Table 3.2**) were identified, each of which had unique start and end positions and was detected in at least one patient. The sizes of these distinct variants ranged from 359 to 956,373 bps with a mean length of ~35 kb. The average number of distinct variants per patient was 140 (**Figure 3.2**). CNVs and deletion variants constituted ~90% and 81% of the variants, respectively. About 83% of the distinct variants were rare, occurring in less than 5% of the patients, whereas ~17% of the variants were common occurring in at least 5% of the study cohort. Additionally, the majority of the variants (83.3%) had two CN state (i.e. bi-allelic), while the rest were multi-allelic (**Table 3.2**). Overall, distinct variants were located within 1,527 different CNVRs.

Table 3.2. The main features of the distinct, high-confidence INDELs/CNVs identified in the study cohort.

Variable	Number	
Total number of distinct INDELs/CNVs	3,486	
Mean distinct INDEL/CNV length	35,187 bps	
Length	**Number**	**%**
INDELs	360	10.33
CNVs	3,126	89.67
Frequency	**Number**	**%**
Rare INDELs/CNVs (< 5% of the patients)	2,891	82.93
Common INDELs/CNVs (≥ 5% of the patients)	595	17.07
***Number of INDELs/CNVs per CN state**	**Number**	**%**
INDELs/CNVs with two CN states	**2,905**	**83.33**
(CN= 0) Two copy deletion	685	19.65
(CN= 1) One copy deletion	1,596	45.78
(CN= 3) One copy duplication	607	17.41
(CN= 4) Two or more copy duplication	17	0.49
INDELs/CNVs with multiple CN states	**581**	**16.67**

A. INDELs/CNVs with three CN states	**577**	**16.55**
CN= 0 or 1	543	15.58
CN= 0 or 3	7	0.20
CN= 0 or 4	2	0.06
CN= 1 or 3	13	0.37
CN= 3 or 4	12	0.34
B. Four INDELs/CNVs with four CN states	**4**	**0.12**
CN= 0, 3 or 4	1	0.03
CN= 0, 1 or 4	1	0.03
CN= 0, 1 or 3	2	0.06

CN: Copy number state. CNV: copy number variation; INDEL: insertion/deletion. *The “normal” CN state of 2 copies is not shown.

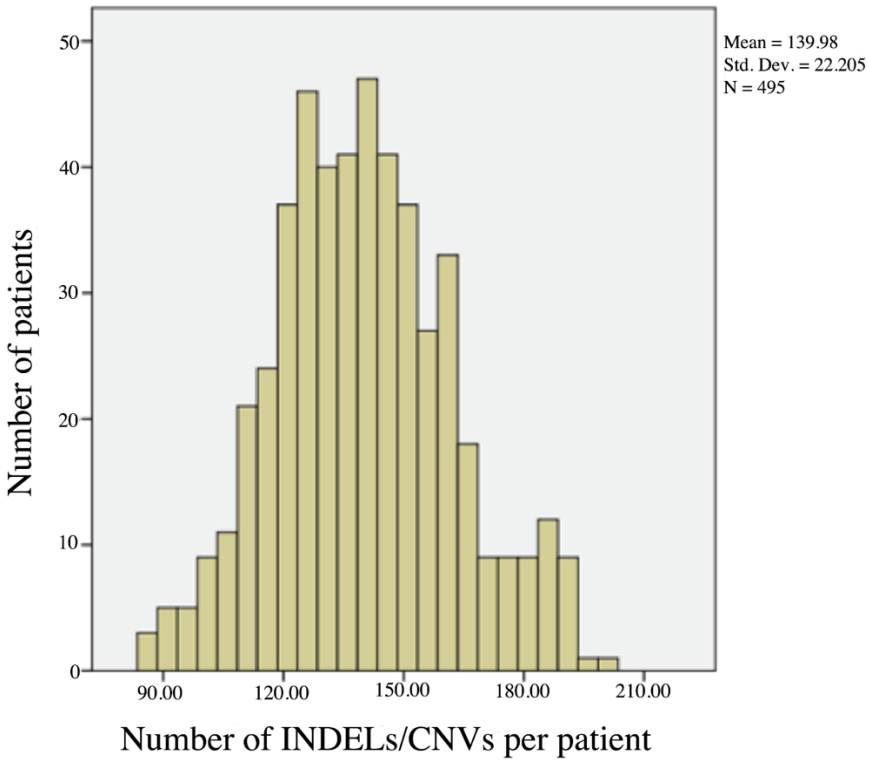


Figure 3.2. Distribution of the number of predicted INDELs/CNVs in the patient cohort. CNV: copy number variation; INDEL: insertion/deletion.

3.6.2 Genes and pathways that may be affected by the distinct INDELs/CNVs

Out of 3,486 distinct INDELs/CNVs, 2,209 (63.4%) variants overlapped with the sequences of 1,673 genes (**Table 3.3**). The entire sequence of 793 genes overlapped with the sequence of a variant; these variants thus may change the gene dosage and affect the transcript levels. A total of 134 genes were affected by multiple INDELs/CNV, representing possible hot-spots. Frequencies of the INDELs/CNVs changed between 0.2% and 45.1% in the patient cohort. The PANTHER database returned information for 742 genes acting in 241 biological pathways.

The main protein pathways that contained the genes affected by the variants are depicted in **Figure 3.3**.

Table 3.3. Genes possibly affected by the INDELs/CNVs.

Affected genes	Numbers
Genes completely covered by INDELs/CNVs	659
Genes partially overlapped with INDELs/CNVs	880
Genes completely or partially overlapped with different INDELs/CNVs	134

CNV: copy number variation; INDEL: insertion/deletion.

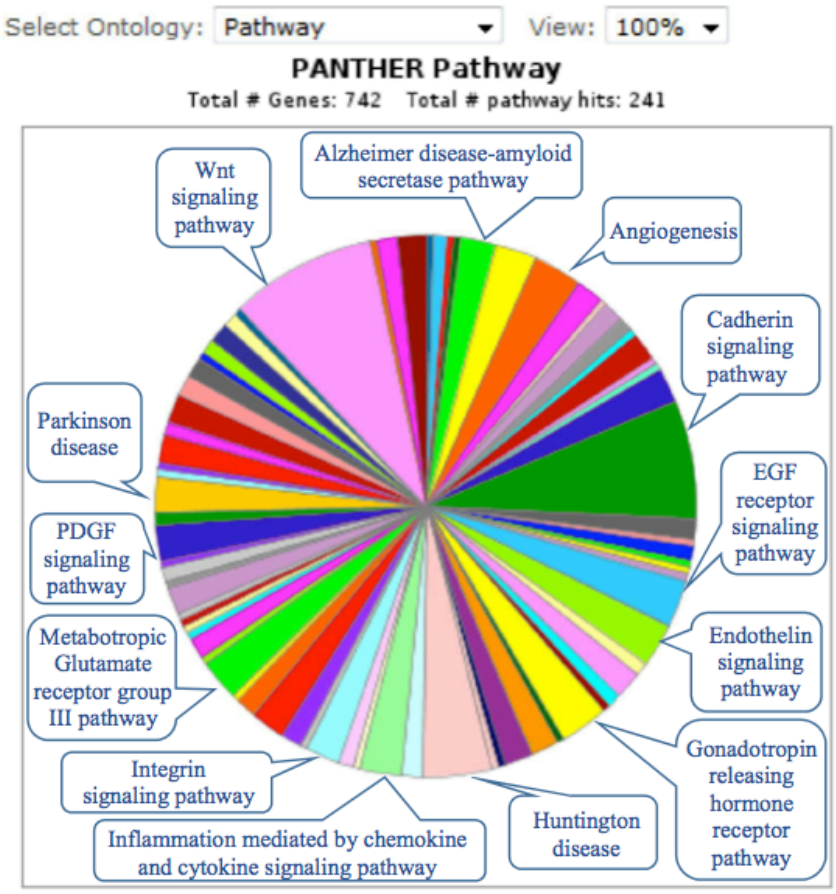


Figure 3.3. PANTHER database results showing the major biological pathways possibly affected by the INDELs/CNVs. CNV: copy number variation; INDEL: insertion/deletion.

3.6.3 DNA analysis

Duplex PCR analysis showed that the results of the computational and experimental analyses agreed in 93% – 100% of the cases (**Appendix C**). Specifically, in the majority of the cases (n=7) the concordance rates were 100%, while in three variants we obtained concordance rates of 99%, 98% and 93%. The lowest concordance rate (93%) was observed in the case of a CNV located in a duplicated gene region (*LCE3C/LCE3B*).

3.6.4 INDELs/CNVs in FCCX cases

There were 13 FCCX cases in our patient cohort. In order to explore whether there were INDELs/CNVs unique/specific to these patients, we first compared the unique and high-confidence variant data of the 13 patients with the rest of the patients in our cohort. As a result, we have identified 28 variants in 11 FCCX patients that were unique to the FCCX cases (**Supplementary Tables 9** and **10**). Twenty-one of these variants affected at least one gene and none of the CNVs or the genes were detected in more than one patient. However, there were two patients who had different variants at chromosome 6p22.1 that overlapped with each other (**Supplementary Table 9**). Second, considering the possibility that rare variants that may be specific to FCCX cases could have been eliminated during the quality control analyses (particularly when we have filtered out the variants that were not detected in previous studies [357–359]), we also looked at the variant data of FCCX cases eliminated at this stage. As a result, there were 30 variants (25 affecting at least one gene) in 13 FCCX cases, which were not identified in other patients in our cohort or the individuals in three other previous studies (**Supplementary Table 10**).

3.6.5 Examination of INDELs/CNVs in relation to relapse-free survival of patients [ii]

Assuming that the PH assumption held for all variables, our results showed that two CNVs (located within the introns of *TGFBR3*, and *STEAP2-AS1 & STEAP2* genes) and one INDEL (located within the intron sequences of the *CMSS1 & FILIP1L* genes) were associated with the relapse-free survival time when adjusted for prognostic factors (**Supplementary Table 7**). In the case of the *CMSS1 & FILIP1L* INDEL, patients with homozygous deletion had increased risk of relapse compared to patients with at least one copy, where as those patients having homozygous deletion of the *TGFBR3* or *STEAP2-AS1 & STEAP2* CNV sequences had reduced risk of relapse compared to patients who had no homozygous deletion of these variants.

We then checked the PH assumption starting with the baseline variables and found that age at diagnosis had time-varying associations with RFS; patients who were younger than 65 were at significantly increased risk of recurrence, metastasis, or death in the initial 2.1 years relatively to the patients who were 65 or older at the time of diagnosis whereas after this time period, the direction of the effect was reversed (i.e. HR: 0.44, p=0.006 and HR: 1.6, p=0.0075, respectively). Thus, we re-analyzed the associations of the variants in age-stratified multivariable models. These analyses identified three variants that have potential time-varying associations with relapse-free survival (**Table 3.4**). Associations of two of these variants with the relapse-free survival time remained significant prior to their time-points where the effect on the relapse-free survival changed (around 3 years post-diagnosis; **Table 3.4**). These CNVs were located within

[ii] The results shown in this section were generated based on work done by the thesis author (i.e., Yajun Yu).

the *PDLIM3* and *GUSBP1* genes and patients with the homozygous deletions of these CNVs had increased and decreased risk of relapse during the initial years after diagnosis, respectively. In the case of the remaining 103 variants that satisfied the PH assumption, in addition to *TGFBR3*, *STEAP2-AS1 & STEAP2*, and *CMSS1 & FILIP1L* variants, association of a new variant overlapping with the sequence of the *RP11-143P4.2* gene was detected in age-stratified models (**Table 3.5**; **Supplementary Table 8**). All of these CNVs/INDELs were located within the intron sequences of the genes.

Table 3.4. Results of the Cox regression models with time-varying coefficients for the three variants that violated the proportionality assumption.

Time-point (years)	Variables in the model	HR	95% CI for HR (lower)	95% CI for HR (higher)	p value	p-value for PH assumption test
4.3	Stage (II vs. I)	1.433	0.856	2.398	0.171	0.588
	Stage (III vs. I)	2.266	1.374	3.736	**0.001**	0.568
	Stage (IV vs. I)	5.950	3.441	10.289	**1.74E-10**	0.146
	Location (Rectum vs. colon)	1.411	1.046	1.904	**0.024**	0.111
	MSI status (MSI-H vs. MSS/MSI-L)	0.327	0.152	0.708	**0.005**	0.230
	*Chr1_169207360_169241309 (0 CN vs. 1 or 2 CN) (*NME7*)					
	Before the time-point	1.400	0.848	2.310	0.188	0.906
	After the time-point	0.159	0.022	1.153	0.069	0.898
2.6	Stage (II vs. I)	1.502	0.899	2.509	0.120	0.832
	Stage (III vs. I)	2.390	1.450	3.940	**0.001**	0.800
	Stage (IV vs. I)	6.591	3.807	11.412	**1.65E-11**	0.082
	Location (Rectum vs. colon)	1.419	1.051	1.916	**0.022**	0.183
	MSI status (MSI-H vs. MSS/MSI-L)	0.315	0.145	0.683	**0.003**	0.206
	*Chr4_186441932_186444110 (0 CN vs. 2 CN) (*PDLIM3*)					

	Before the time-point	2.108	1.317	3.373	**0.002**	0.794
	After the time-point	0.726	0.423	1.245	0.244	0.864
2.8	Stage (II vs. I)	1.477	0.883	2.470	0.138	0.678
	Stage (III vs. I)	2.354	1.428	3.879	**0.001**	0.693
	Stage (IV vs. I)	5.952	3.448	10.274	**1.52E-10**	0.086
	Location (Rectum vs. colon)	1.421	1.052	1.919	**0.022**	0.103
	MSI status (MSI-H vs. MSS/MSI-L)	0.323	0.149	0.700	**0.004**	0.224
	*Chr5_21450792_21452439 (0 CN vs. 2 CN) (*GUSBP1*)					
	Before the time-point	0.416	0.182	0.955	**0.039**	0.770
	After the time-point	1.511	0.927	2.463	0.098	0.848

Chr: chromosome; CI: confidence interval; CN: copy number state; HR: hazard ratio; MSI-H: microsatellite instability-high; MSI-L: microsatellite instability-low; MSS: microsatellite stable; PH: proportional hazards; vs.: versus. P-values < 0.05 are bolded. *Genes that overlap with the variants are shown in parentheses.

Table 3.5. Variants that satisfied the proportionality assumption and significantly associated with the relapse-free survival time.

Gene	Variant	p-value	HR	95% CI (lower)	95% CI (higher)
TGFBR3	Chr_1_92232111_92233227 (0 CN vs 2 CN)	0.0454	0.5211	0.2752	0.9867
CMSS1, FILIP1L	Chr_3_99628822_99629567 (0 CN vs 1 or 2 CN)	0.015	1.6936	1.1076	2.5896
RP11-143P4.2	Chr_3_192875738_192885153 (0 CN vs 2 or 4 CN)	0.0394	1.3586	1.0149	1.8186
STEAP2-AS1, STEAP2	Chr_7_89810608_89812114 (0 CN vs 2 CN)	0.0372	0.5776	0.3447	0.968

Chr: chromosome; CI: confidence interval; CN: copy number state; HR: hazard ratio.

3.7 Discussion

In this study we detected the genome-wide INDEL/CNV profiles of 495 Caucasian colorectal cancer patients from Newfoundland, Canada, using two CNV detecting algorithms and stringent quality control measures. Further analyses were performed to test the associations of 106 genic and common variants with the patient outcomes. The potential time-varying associations of these variants on relapse-free survival times were also investigated. Additionally, we explored the rare and unique INDELs/CNVs that are only observed in 13 hereditary colon cancer syndrome patients diagnosed with FCCX.

Our results showed that, similar to other studies QuantiSNP and PennCNV detected different numbers of variants in the patient genomes, which can be attributed to the different methodologies applied by these algorithms [360,361]. However, when a variant was detected by both algorithms, the genomic positions and borders of the variants were identical in the majority of the cases (84.3%), suggesting a high-concordance rate for variants detected by both QuantiSNP and PennCNV. In addition, 97% of the variants after the quality control measures had at least 50% of their sequences overlap with the variants previously identified by other groups. These results are in agreement with others' findings [360–362] that the false-prediction rate decreases when multiple algorithms and strict quality control measures are used for INDEL/CNV detection. This was further supported by the DNA analysis of 10 of the variants in our study, which showed a fairly high concordance rate between the DNA analyses and the computational predictions.

The majority of the variants identified in this study were deletions (**Table 3.2**). This is expected as when a genome-wide signal intensity data is used, deletion variants are detected easier than duplication variants (CN ≥ 3) [353]. Also, our list of variants contain mostly the large

variants (i.e. CNVs with sizes of at least 1 kb). This too is expected because the QC measures inclined towards removing smaller variants. For example, during this study variants with sizes < 10 bps or detected by < 10 probes were eliminated from the variant calls to remove the potential false-positives. These criteria inevitably should have resulted in exclusion of a portion of the short variants. Of note, the shortest high-confidence variant identified in our study had a length of 359 bps. Therefore, while it is likely that our variant data is missing a portion of variants due to the strict QC measures, our QC measures also served to reduce the false-positive predictions, increased the accuracy of our results, and at the end yielded INDELs/CNVs that are deemed to be detected with high-confidence.

The sequences of a number of variants we identified overlap with the human gene sequences. These "genic" INDELs/CNVs are biologically interesting as they can delete or duplicate gene sequences, and as a result may affect physiological functions. Overall, our data showed that the number of gene sequences affected by rare variants (n=1,538) were higher than the number of gene sequences affected by common variants (n=135). Similar to others' findings, these results may be explained by the fact that variants that affect genes are kept at low frequencies in the populations [363]. Additionally, the genes that harbour INDEL/CNV sequences come from a variety of biological pathways (**Figure 3.3**), some of which are established in cancer development or progression; notably WNT signaling and angiogenesis pathways [364–367]. Variants identified in this study hence deserve further investigation as it is possible that some of them are biologically linked to susceptibility or prognosis in colorectal cancer.

Considering that rare INDELs/CNVs may lead to high-penetrant genetic disorders including FCCX, as part of this study we also explored the variant data in 13 FCCX cases. FCCX is a familial colon cancer syndrome where patients satisfy the clinical criteria for

hereditary non-polyposis colorectal cancer (HNPCC) but have tumors that lack the microsatellite instability [368]. Many different genetic approaches including linkage, association, CNV, and mutation screening studies, have been performed in FCCX cases/families. While these studies have identified several candidate genes and genetic regions, the entire body of findings suggest genetic heterogeneity and lack of a common genetic cause among unrelated FCCX cases [347,369–371]. In this study we have examined the INDEL/CNV profiles of the FCCX cases in our cohort and identified a number of rare variants that were unique to the FCCX patients. Our results, however, did not identify a gene or INDEL/CNV that was detected in multiple unrelated cases (although we have identified two patients with overlapping variants on chromosome 6p22.1). Thus, our data largely agree with previous findings and do not provide an evidence of specific rare variants or genes that can explain this disease in more than one FCCX patients. We also compared our findings with the others in the literature. A study by Masson et al [347] suggested the involvement of CNVs, at least to some extent, in FCCX development. A comparison of the INDELs/CNVs only detected in our FCCX patients (**Supplementary Tables 9** and **10**) and Masson's group did not identify a common variant or gene affected by the variants in our list. However, there were a number of CNVs/INDELs in our data that were located within or around the genomic regions previously identified in linkage analyses (summarized in Sanchez-Tome et al. 2015 [370]). These INDELs/CNVs thus may form an interesting list of candidate variants for further studies that can dissect the potential INDEL/CNV – FCCX relationship.

Considering the fact that colorectal cancer patients have increased risk of death as well as recurrence and metastasis after their initial diagnosis/treatment [5,337,372], we also examined the associations of baseline clinical factors and 106 CNVs/INDELs with the survival outcome in our patient cohort. We note that while the results obtained are generally quite similar, since it is the

proper model for variants that violate the proportionality assumption, we consider the results of the Cox regression model with time-varying coefficients (**Table 3.4**) more accurate than the results of the conventional Cox PH regression model. One of the interesting findings of this analysis was that the hazards ratio of age at diagnosis categories (< 65 years versus ≥ 65 years) changed over time. Specifically, relatively young age at diagnosis (< 65 years) was associated with increased risk of relapse within the first ~2 years after diagnosis, while after this initial time period the risk of relapse increased for the older patients (≥ 65 years). The exact reason of this time-varying association in our patient cohort is not known, but it can be linked to aggressive or advanced disease at diagnosis in relatively younger patients in our cohort (46.8% stage III and IV patients in < 65 years of age category compared to 36.6% stage III and IV patients in the ≥ 65 years of age category). Although different criteria are used for young patient classification in other studies (which is usually < 40 years of age [147,373–375]), this is consistent with the other published reports where the younger patients were reported to be more likely to be diagnosed at later stages and have increased chance of recurrence early after diagnosis [372,376].

As per the genetic variants, our analyses identified a total of six genic variants (five CNVs and one INDEL) that were associated with the relapse-free survival time in the patient cohort (**Tables 3.4** and **3.5**). The sizes of these variants changed from 746 bp – 9,416 bp and all were located in non-coding (i.e. intronic) parts of the genes. The genes that may be affected by these variants function in a variety of biological pathways; *PDLIM3* codes for a cytoskeletal protein; *GUSBP1* codes for an expressed pseudogene with unknown functions; *TGFBR3* codes for a TGFβ signaling pathway protein; *STEAP2-AS1* codes for the antisense RNA for *STEAP2* and *STEAP2* codes for a transmembrane metalloreductase; *RP11-143P4.2* codes for a long non-coding RNA; and *CMSS1* codes for a ribosomal small subunit homolog and *FILIP1L* codes for a

filamin A-binding like protein. Some of these genes were previously linked to carcinogenesis and disease progression. For example, *TGFBR3* is a potential tumor suppressor gene deleted in various cancers and with a role also in cell migration, invasion, and metastasis [377]. Interestingly, one study reported its expression being associated with reduced apoptosis and increased migration in a colon cancer cell line [378]. Additionally, FILIP1L has been shown to have a role in inhibition of WNT signaling pathway, a pathway implicated in colorectal cancer and metastasis [363,367] as well as in cellular invasion in an ovarian cancer model [379] and colon cancer cell lines [380]. Consistent with these results, another study showed that reduced levels of this protein in colorectal tumors were associated with reduced overall survival times of patients [381]. While it is currently unknown whether these INDELs/CNVs have biological effects on the corresponding genes (and hence, have direct effects on the disease progression and risk of relapse in colorectal cancer), it is quite possible as a large number of non-coding sequences in the human genome contain regulatory elements [382].

Literature search showed that none of these six variants were previously linked to outcome in colorectal cancer patients, or patients diagnosed with other cancers. Interestingly, we identified that the relationships of two of these variants with the risk of relapse have varied with time (**Table 3.4**). Specifically, the hazard ratios by the *GUSBP1* and *PDLIM3* CNVs fluctuated over time, with a statistically significant associations detected only early after diagnosis (i.e. within the first ~3 years), but not after these years. Both of these CNVs are common variants presenting in 14% and 20% of the patient cohort (*GUSBP1* and *PDLIM3* CNVs, respectively). These results may be explained by these genetic variants either directly and biologically affecting the risk of recurrence/metastasis, or death, or being correlated with a yet unknown factor(s) that modifies the risk of relapse during this time period. We also note that their

associations were detected only when the statistical analyses considered the time-varying associations; otherwise these associations were missed when conventional Cox regression method was used (**Supplementary Tables 7** and **8**). This highlights the importance of using appropriate statistical approaches that can help uncover novel findings that are otherwise prone to be missed. Currently, examining the potential time-varying associations of genetic polymorphisms/mutations on the risk of outcome is quite a rare practice. To our knowledge, previously only one study has examined and identified a genetic marker with a possible time-varying association with the risk of outcome in colorectal cancer. In short, Pavelitz et al [196] examined the *MRE11* gene mutation status in stage III colorectal cancer patients and found that the proportionality assumption of the Cox modeling was violated for overall and disease free survival times in their patient cohort. These authors then moved on with a statistical approach that we adapted in our analysis, including identification of a time-point and modeling survival outcome using the Cox regression model with time-varying coefficients [383]. Therefore, the mutant *MRE11* these authors identified and the germline *GUSBP1* and *PDLIM3* CNVs our study identified are the first examples of genetic markers that potentially have time-varying associations with patient outcomes in colorectal cancer. Overall, we conclude that the *GUSBP1* and *PDLIM3* CNVs are potential early-relapse markers in colorectal cancer, and if results obtained in this study are replicated they can be useful not only in developing more informative prognostic models but also in elucidating the biological basis of variable risk of relapse (i.e. risk of recurrence, metastasis, or death) among colorectal cancer patients.

Like other studies, this one has strengths and limitations. Our main strengths were the followings; **a)** the Illumina® Omni-1-quad platform used to generate the genome-wide signal intensity data and helped detection of INDELs/CNVs is a high-resolution platform, which

facilitates a more efficient variant detection compared to many other platforms; **b)** two CNV detection algorithms and stringent quality control/filtering steps were used in order to reduce the false-positive predictions; **c)** the results of the computational INDEL/CNV detection and the duplex PCR analysis were largely condordant; **d)** this is the first large-scale analysis of germline genic INDELs/CNVs and their relation to relapse-free survival in colorectal cancer; **e)** this is the first study that identified germline polymorphisms with time-varying associations with patient outcome in colorectal cancer; and **f)** the patient cohort was a well-described cohort with a long follow-up time, which increased our study power. Our limitations were; **a)** variants from sex chromosomes were not included in the computational analyses; **b)** while our approach detected INDELs, a significant portion of the INDELs remained unidentified as the detection parameters were geared towards detection of larger variants; **c)** rare variants were not examined in relation to survival outcomes; **d)** the experimental analyses were limited to duplex PCR assessing the homozygous deletion and copy number states ≥ 1 rather than quantitative techniques that could detect the individual copy number states; **e)** the patient cohort was of Caucasian ancestary, thus the results may not be applicable to patients from other populations.

In conclusion, this is one of the first studies that identified the genome-wide INDEL and CNV profiles in a large cohort of colorectal cancer patients. Our variant data is in line with the results of other studies reported in the literature. This is also the first study that comprehensively investigated the possible associations of genic INDELs/CNVs with relapse-free survival time in colorectal cancer. We identified six variants that are candidate prognostic markers and should be examined in futher studies. This is also the first study that examined and identified two CNVs that have time-varying associations with clinical outcomes of colorectal cancer patients; if replicated these CNVs can be used as early-relapse markers during prognostication. Last but not

the least, this study suggests that similar to other literature findings there was no one, unique, and rare INDEL or CNV that could explain the risk of FCCX in unrelated patients. Overall, this study has important implications for the future studies of INDELs/CNVs and susceptibility and prognosis in colorectal cancer.

3.8 Acknowledgements

The study team is indebted to the patients who participated in the Newfoundland Familial Colorectal Cancer Registry (NFCCR) for making this study possible and many investigators and staff for their contributions to the NFCCR over the many years. We gratefully acknowledge Dr. Roger Green (RG), who contributed to both NFCCR and the initial steps of this study, who sadly passed away prior to completion of this project. We also thank Ms. Michelle Simms for her help with the patient DNA samples, and Dr. Kai Wang (at University of Pennsylvania) for patiently and promptly responding to our queries regarding the CNV predictions using the PennCNV algorithm.

This study was funded mainly by Colon Cancer Canada (funds to SS) and The Dean's Innovation Fund – The Medical Research Foundation at Faculty of Medicine, Memorial University (funds to SS and YEY). NFCCR and generation of the genomewide SNP genotype data were previously funded by other granting agencies including the Research and Development Corporation of Newfoundland (RDC; leverage fund to WX, RG, PP, SS: contract number: 5404.1201.102), Canadian Institutes of Health Research (CIHR; RPP-operating funds to WX, RG, PP, SS; FRN: 110045), Medical Research Fund (MRF) of Memorial University (funds to SS and RG), CIHR fund for the Colorectal Cancer Interdisciplinary Health Research Team at the

University of Toronto and Memorial University (awarded to the NFCCR and other investigators), the National Cancer Institute of Canada (awarded to the NFCCR investigators) and the Atlantic Innovation Fund for the Interdisciplinary Research Team in Human Genetics (awarded to the NFCCR investigators). YY is supported by a Translational and Personalized Medicine Initiative (TPMI) Educational Funding fellowship. The funding sources had no involvement 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. SS is a Beatrice Hunter Cancer Research Institute (BHCRI) senior scientist.

CHAPTER 4: The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying associations

A version of this manuscript has been published in BMC Medicine; 2019, 17(1):150. The manuscript in this Chapter had only minor changes compared to the published version (e.g., "time-varying effects" was changed to "time-varying associations". This is for keeping consistency of terminology throughout the thesis). Note that supplementary information that was published with the manuscript is presented in Appendix D.

Yajun Yu[1], Megan Carey[1], William Pollett[2#], Jane Green[1], Elizabeth Dicks[3], Patrick Parfrey[3], Yildiz E. Yilmaz[1,3,4], Sevtap Savas[1,5]

[1] Discipline of Genetics (As of Sep 2020, the Discipline of Genetics has become a part of the Division of Biomedical Sciences), Faculty of Medicine, Memorial University, St. John's, NL, Canada.

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

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

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

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

[#]Retired

4.1 Co-authorship statement

Yajun Yu helped design the statistical approach, performed all analyses, interpreted the results, and drafted the manuscript.

Megan Carey helped with the collection of patient-related data.

William Pollett helped with the collection of patient-related data.

Jane Green helped with the collection of patient-related data.

Elizabeth Dicks helped with the collection of patient-related data.

Patrick Parfrey helped with the collection of patient-related data.

Yildiz E. Yilmaz helped with the statistical methodology.

Sevtap Savas conceptualized and led the study; helped with the collection of patient-related data; helped design the study and interpret the results, and revised and submitted the manuscript.

4.2 Abstract

Background: Colorectal cancer is the third most common cancer in the world. In this study, we assessed the long-term survival characteristics and prognostic associations and potential time-varying associations of clinico-demographic variables and two molecular markers (microsatellite instability [MSI] and BRAF Val600Glu mutation) in a population-based patient cohort followed up to ~19 years.

Methods: Patient cohort included 738 incident cases diagnosed between 1999 and 2003. Cox models were used to analyze the association between the variables and a set of survival outcome measures (Overall survival: OS, Disease-specific survival: DSS, Recurrence-free survival: RFS, Metastasis-free survival: MFS, Recurrence/Metastasis-free survival: RMFS, and Event-free survival: EFS). Cox proportional hazard (PH) assumption was tested for all variables, and Cox models with time-varying associations were used if any departure from the PH assumption was detected.

Results: During the follow-up, ~61% patients died from any cause, ~26% died from colorectal cancer, and ~10% and ~20% experienced recurrences and distant metastases, respectively. Stage IV disease and post-diagnostic recurrence or metastasis were strongly linked to risk of death from colorectal cancer. If a patient had survived the first 6 years without any disease-related event (i.e. recurrence, metastasis, or death from colorectal cancer), their risks became very minimal after this time period. Distinct sets of markers were associated with different outcome measures. In some cases, effects by variables were constant throughout the follow-up. For example, MSI-high tumor phenotype and older age at diagnosis predicted longer MFS times consistently over the follow-up. However, in some other cases, effects of the variables varied

with time. For example, adjuvant radiotherapy treatment was associated with increased risk of metastasis in patients who received this treatment after 5.5 years post-diagnosis, but not before that.

Conclusions: This study describes the long-term survival characteristics of a prospective cohort of colorectal cancer patients, relationships between baseline variables and a detailed set of patient outcomes over a long time, and time-varying associations of a group of variables. Results presented advance our understanding of the long-term prognostic characteristics in colorectal cancer and are expected to inspire future studies and clinical care strategies.

4.3 Background

Colorectal cancer is an important disease to control. It is one of the most commonly diagnosed cancers in the world, causing ~700,000 deaths each year [2]. Many patients with colorectal cancer also experience clinically important events, such as recurrences or metastases after diagnosis. Assessing the characteristics of potential disease outcomes and identifying their predictors are critical for effective patient surveillance, and to treat and control this disease in both the short-term and long-term. Studies have reported that the majority of the recurrences, metastases, and deaths from colorectal cancer occur within the first few years following the diagnosis or surgery [384,385]. The main clinical surveillance guidelines recommend up to 5 years of follow-up [79].

Clinical features (e.g. disease stage, tumor grade, histology, location), demographic variables (e.g. age at diagnosis, sex, and familial risk status), and tumor characteristics (e.g. the MSI tumor phenotype and somatic mutations, including BRAF Val600Glu mutation) are among the most commonly investigated variables in colorectal cancer [137,138,386–389]. Familial risk status may indicate familial clustering of the disease and is an interest for both the susceptibility and prognostic studies [389,390]. Microsatellite instability (MSI) is a tumor phenotype that is characterized by defects in the DNA mismatch repair system that lead to the genomic instability [391]. Generally, MSI-high tumors are associated with better patient survival [387]. BRAF Val600Glu mutation occurs in ~10% of the colorectal tumors, causes oncogenic BRAF activity, and promotes cellular transformation [388]. Literature reports also suggest a prognostic role for this BRAF mutation in colorectal cancer [392].

Many prognostic studies aim to identify markers to help distinguish the patients with different outcome risks. Potential time-varying associations of markers on the patient outcomes, however, are not well-studied. Markers with time-varying associations are those whose effect direction (e.g. protective or detrimental) or size (i.e. magnitude) changes over the follow-up [185,191–193,393]. There are at least two important implications of assessing the time-varying associations of the markers in prognostic studies. First, such markers are important as they can distinguish the patients who are at increased risk of events only during specific time-periods (e.g. in the short-term [early event markers], or the long-term [late event markers or markers with late effects]). Second, examining the time-varying associations of variables is not a standard or widely utilized research practice, which potentially leads to loss of information or inaccurate inference [184,185,394].

In colorectal cancer, a few studies examined the clinical, demographic, or molecular variables for time-varying associations using statistical methods. For example, disease location [190]; age, disease stage, time period of diagnosis, or tumor site [191–193]; regional cancer, age, and tumor location (pelvic/sigmoid colon) [194]; age (in two of our previous studies using subsets of the patients included in this study) [395,396]; tumor site (left or right), grade, sex, and stage [195]; a set of genetic variations [177,395]; and a somatic tumor alteration [196] were reported to have or tend to have time-varying associations with patient outcomes. Among the statistical methods that are used for identification of time-varying associations are the mixture cure model [397], Cox-Aalen model, additive models with time-varying associations, and multiplicative models with time-varying associations (including piece-wise/change-point Cox proportional hazards [PH] regression model) [210,383]. Cox PH regression model [199,200] is one of the most widely used statistical model for time-to-event analyses in medical sciences [184]. This model has an assumption (the PH assumption) where the hazard ratio for any two groups of patients stratified by a variable remains constant over time. Violation of the PH assumption implies that the effect of the variable being investigated changes over-time [199,200,218]. Hence, assessing the PH assumption in Cox models is an opportunity to identify the variables that have time-varying associations with patient outcomes.

While colorectal cancer is a common disease in the world, long-term prognostic characteristics and their predictors are not well known. In this study, we investigated the data collected from a prospective colorectal cancer patient cohort followed up to 19 years. Our specific aims were to examine: 1) the long-term survival characteristics; and 2) associations as well as the potential time-varying associations of the widely investigated baseline clinico-

demographic variables and select molecular markers on a comprehensive set of patient outcome measures.

4.4 Methods

4.4.1 Patient cohort, patient-related data, and inclusion criteria

This is an observational study. The patient cohort examined in this study was recruited by the Newfoundland Colorectal Cancer Registry [328,329]. This registry includes 750 incident cases, who were diagnosed with colorectal cancer in Newfoundland and Labrador (NL) between January 1999 and December 2003. These patients constituted 64% of the eligible patients who were diagnosed within this time frame. Among the 750 patients recruited by the registry, clinical and prognostic data of 744 patients were available in the registry and were provided to the study team. Out of 744, 738 colorectal cancer patients with stage I-IV disease and an age at diagnosis ≤ 75 were included in the present study (5 patients with in situ/stage 0 tumors and one patient with >75 years of age were excluded). In this study, clinical, pathological, demographic, and molecular markers that are most widely examined by the colorectal cancer research community and present in at least 5% of the patient cohort were selected for assessment (**Table 4.1**). Tumor MSI and BRAF Val600Glu mutation statuses were determined previously as described in Woods et al. [329], and familial risk status was assessed as described in Green et al. [328]. Information on the clinical, pathological, and demographic as well as the vital status, cause of death, recurrence, and metastasis was collected over time using several resources as described in Negandhi et al. [262] that

included patient follow-up questionnaires, medical records (e.g. physician notes/assessments, pathology, surgery, and autopsy reports/death certificates), Provincial Tumor Registry-NL/Dr. H. Bliss Murphy Cancer Centre, and NLCHI (Newfoundland and Labrador Centre for Health Informatics). The distinction between loco-regional recurrence and distant metastasis was based on the pathology reports, diagnostic imaging reports, location of tumors, or physician's notes. If a tumor had occurred in the field of the primary resected tumor, including proximal or distal to the site of anastomosis, it was classified as recurrence. Distant recurrences were classified as metastasis based on the location and clinical assessment of the origin of the tumor, and physicians' opinions.

The last date of follow-up in this cohort was January 2018. An overview of the characteristics of the clinico-demographic variables and molecular markers of the patient cohort is shown in **Table 4.1**.

Table 4.1. Baseline characteristics of the patient cohort.

Variable	Number	Percentage (%)
Age at diagnosis		
Median (range)	62.37 (20.70 - 75.01)	-
Sex		
Male	452	61.25
Female	286	38.75
Familial risk		
Low risk	355	48.10
High/intermediate risk	362	49.05

Unknown	21	2.85
Location		
Colon	507	68.70
Rectum	231	31.30
Stage		
I	113	15.31
II	245	33.20
III	227	30.76
IV	153	20.73
Histology		
Non-mucinous	646	87.54
Mucinous	92	12.47
MSI status		
MSI-L/MSS	636	86.18
MSI-H	73	9.89
Unknown	29	3.93
BRAF Val600Glu mutation status		
Wild-type	591	80.08
Mutant	80	10.84
Unknown	67	9.08
Grade		
Well/moderately differentiated	653	88.48
Poorly differentiated	73	9.89
Unknown	12	1.63
Adjuvant chemotherapy treatment		
No	387	52.44
Yes	346	46.88

Unknown	5	0.68
Adjuvant radiotherapy treatment		
No	565	76.56
Yes	151	20.46
Unknown	22	2.98

MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable.

4.4.2 Statistical analyses

4.4.2.1 Assessing the collinearity among the variables

We assessed and ruled out the potential correlation between the categorical variables (**Table 4.1**) based on the pair-wise Pearson's correlation coefficient value (**Supplementary Table 11**).

4.4.2.2 Survival outcomes

A set of widely-investigated survival outcomes were examined in order to conduct a comprehensive investigation. The endpoints of overall survival (OS), disease-specific survival (DSS), recurrence-free survival (RFS), metastasis-free survival (MFS), recurrence or metastasis-free survival (RMFS), and event-free survival (EFS) were death from any cause, death from colorectal cancer, diagnosis of local recurrence, diagnosis of distant metastasis, diagnosis of recurrence or metastasis, and diagnosis of recurrence, metastasis, or death from colorectal cancer, respectively. Survival times were calculated starting at the date of diagnosis till the date

of the first occurrence/observation of the endpoint (or the date of last contact); in multi-event outcomes the latter date was the date of the first event. In each survival outcome, data were censored at the date of last contact for patients who have not experienced the events of interest during their follow-up. Data on the survival outcomes are summarized in **Table 4.2**.

Table 4.2. Number of events in the survival outcomes examined in this study.

Survival status	Number	Percentage (%)
OS status		
Alive	290	39.30
Died	448	60.70
DSS status		
Death from other causes or alive	399	54.07
Death from colorectal cancer	192	26.02
*Unknown	147	19.92
RFS status		
Recurrence (-)	661	89.57
Recurrence (+)	77	10.43
MFS status		
Metastasis (-)	587	79.54
Metastasis (+)	151	20.46
RMFS status		
Recurrence or metastasis (-)	542	73.44
Recurrence or metastasis (+)	196	26.56
EFS status		

Recurrence, metastasis, or death from colorectal cancer (-)	359	48.64
Recurrence, metastasis, or death from colorectal cancer (+)	287	37.67
*Unknown	101	13.69

DSS, disease-specific survival; EFS, event-free survival; MFS, metastasis-free survival; OS, overall survival; RFS, recurrence-free survival; RMFS, recurrence/metastasis-free survival. *This is because the cause of death information was missing for some patients.

4.4.2.3 Kaplan-Meier and Cox regression analyses, and Proportional Hazards (PH) assumption test

IBM SPSS Statistics (version 25) was used for Kaplan-Meier method that generated the survival curves. Univariate Cox models were fitted for variables for each of the survival outcomes. PH assumption test [204] was performed using cox.zph function [205] in R (ver. 3.5.0) [206] using the default "km" function to obtain the transformed survival times. Multivariable Cox models were constructed using backward selection method and when the PH assumption was violated, Cox model with time-varying associations (assuming piece-wise constant hazard ratios – this model is also called change-point Cox model [210,383]) was used. Full multivariable models with all baseline variables were first checked for the PH assumption. For the variables that violated the PH assumption, cut-off time points before and after which the PH assumption was satisfied were obtained, starting from the variable with the lowest p-value of the PH assumption test. The proper cut-off time points were selected based on the approach described in Pavelitz et al. [196] and Klein and Moeschberger [210]. In this study a set of cut-off time points ranging from 0.5 years to 18.5 years and with increments of 0.5 years were considered. The proper cut-off time

point is ideally the one which makes the model (1) having the largest maximized log partial likelihood, and (2) with the PH assumptions being satisfied before and after the time point. If the model with the largest maximized log partial likelihood did not satisfy the proportional hazards at both time intervals separated by the tested time point, then the one with the second largest maximized log partial likelihood value was tested. This step was repeated until a model was obtained that satisfied the criteria. The corresponding cut-off time point was then deemed to be the proper cut-off time point. In cases when the cut-off time points made the model having an infinite upper-limit of the 95% confidence interval (CI) for hazard ratio (HR) of a variable, the cut-off point with the next largest maximized log partial likelihood was considered. This is because infinite limits suggest that a valid effect estimation cannot be made. In addition, rarely a proper cut-off time point for a variable was not identifiable. For example for stage III in the OS analysis, a single time point that satisfy the PH assumption in both time periods (before and after the time point) was not identified. We then introduced additional time points in one of the time periods where the PH assumption was violated. However, this step did not identify any proper time points in this region. In this case, we analyzed this variable with the next one in line (i.e. the next variable with the smallest p-value of the PH assumption test) and tested all the possible combinations of the cut-off time points to identify the proper cut-off points of both variables at the same time. Once the proper cut-off time points were identified and included in the model, variables with p-values ≥ 0.05 in the model were removed one by one, starting with the one with the largest p-value. During this process, if any of the remaining variables violated the PH assumption, cut-off time point was identified/re-identified for this variable based on the method described above, followed by re-fitting of the model. The variables in the final model for each outcome measure reported in this manuscript have a p-value < 0.05 either over the follow-up

time (i.e. variables with no time-varying associations), or in at least one time period defined by the cut-off time points (i.e. variables with time-varying associations).

Age at diagnosis was examined as a continuous variable in this study. A p-value < 0.05 was considered significant. All statistical analyses were performed by R (ver. 3.5.0) [206] or IBM SPSS Statistics (version 25).

4.5 Results

4.5.1 Characteristics of the survival outcomes in the patient cohort

Baseline characteristics of the patients and information on the outcome measures investigated are shown in **Tables 4.1** and **4.2**. The median follow-up time was 9.36 years and with a range of 0.04 years to 19.00 years. Among the 738 colorectal cancer patients, 448 (~61%) died by the end of the follow-up period. The number of deaths caused by colorectal cancer (n=192) accounted for ~43% of all deaths. Stage IV patients had the highest rate of death (death from any cause and death from colorectal cancer were recorded for the 94.8% and 86.9% of the stage IV patients, respectively). In addition, 77 patients (~10%) had experienced at least one local recurrence and 151 individuals (~20%) had experienced at least one metastasis. The majority of the patients diagnosed with recurrence or metastasis were stage II or III patients, whereas ~14% of the patients who experienced recurrences and ~7% of the patients who experienced metastases during their follow-up were stage I patients. Around 27% (n=196) of the patients experienced either recurrence or metastasis. The proportion of patients who had

experienced both recurrence and metastasis was low (n=32 patients; 4%), yet almost half of the patients who had recurrence also had metastasis. Approximately 73% of the patients who were diagnosed with recurrence or metastasis died from colorectal cancer (110 out of 150 patients with complete data on recurrence, metastasis, and cause of death). Of the 448 patients died during the follow-up period, 171 patients experienced recurrence or metastasis before they passed away. Overall, ~38% of the patients had at least one disease-related and clinically important event (i.e. recurrence, metastasis, or death from colorectal cancer) during the follow-up.

4.5.2 Survival patterns over time

Kaplan-Meier curves for the outcome measures examined are shown in **Figure 4.1**. Unlike death from all causes (OS: **Figure 4.1A**), the majority (85%) of the deaths due to colorectal cancer occurred within the 6.2 years post-diagnosis (DSS: **Figure 4.1B**). Similarly, the majority (85%) of the first recurrences and/or metastases were diagnosed within the ~5 years after the disease diagnosis (5.1 years for RFS; 4.9 years for MFS; ~4.5 years for RMFS; **Figures 4.1C-E**). As for the EFS that considers the three most important disease-related events, 85% of the first of any of these events were observed within the first ~4.5 years after diagnosis (**Figure 4.1F**). It is important to note that within this group of patients (i.e. with a positive status of recurrence, metastasis, or death from colorectal cancer), only a small portion of the patients (5%) experienced their first disease-related events after the 6 years following the diagnosis of colorectal cancer.

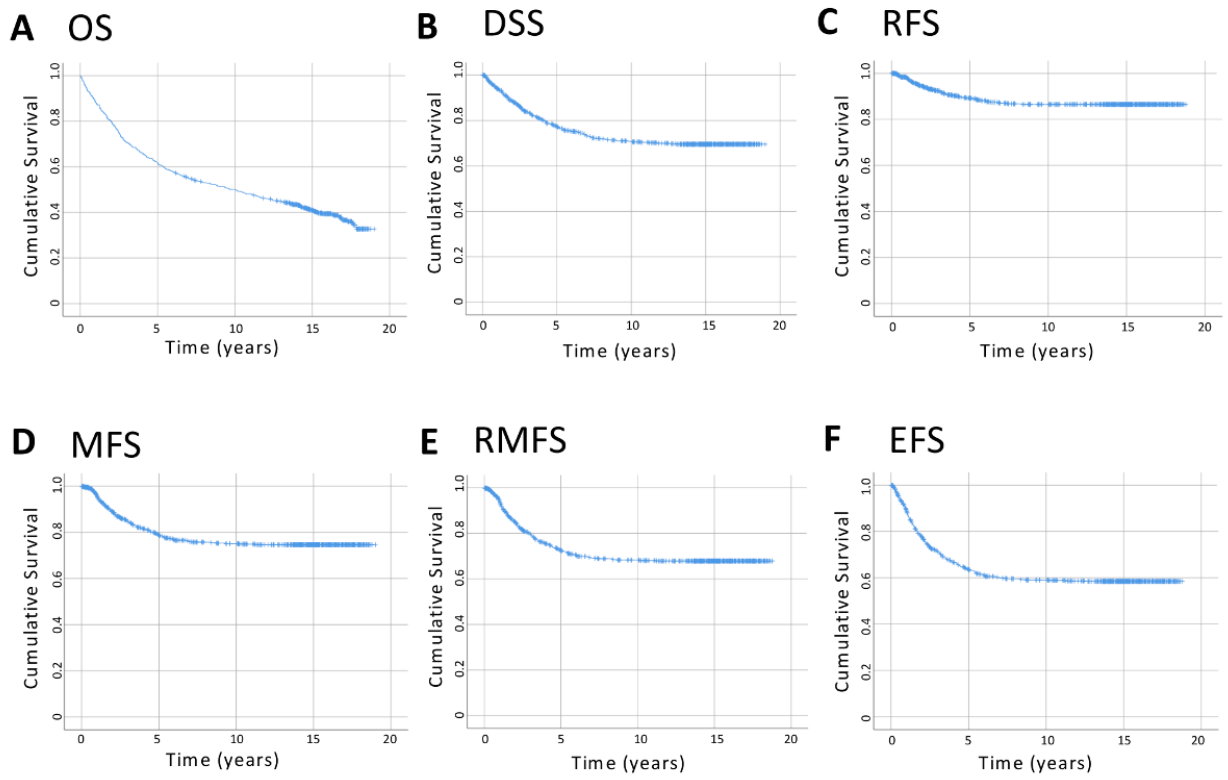


Figure 4.1. **Kaplan-Meier curves of the survival outcomes.** DSS, disease-specific survival; EFS, event-free survival; MFS, metastasis-free survival; OS, overall survival; RFS, recurrence-free survival; RMFS, recurrence/metastasis-free survival.

4.5.3 Variables with or without time-varying associations on survival outcomes

4.5.3.1 Univariate analyses

Univariate associations between clinico-demographic and molecular variables and survival outcomes are summarized in **Figure 4.2**. All variables investigated, except the familial

risk status, were associated with at least one survival outcome. Several variables also violated the PH assumption of the Cox regression models. For these variables, Kaplan-Meier curves showing the survival probabilities over time are depicted in **Supplementary Figures 4-6**. Those variables that violated the PH assumption and were significantly associated with the outcomes (univariate Cox regression p-value < 0.05) tended to have separated curves with no visible crossing of the curves (**Type A variables**; **Supplementary Figures 4** and **6**). In contrast, it was clearly observable that those variables that violated the PH assumption but were not significantly associated with the outcomes (i.e. univariate Cox regression p-value ≥ 0.05) tended to have their curves crossed (**Type B variables**; **Supplementary Figure 5**).

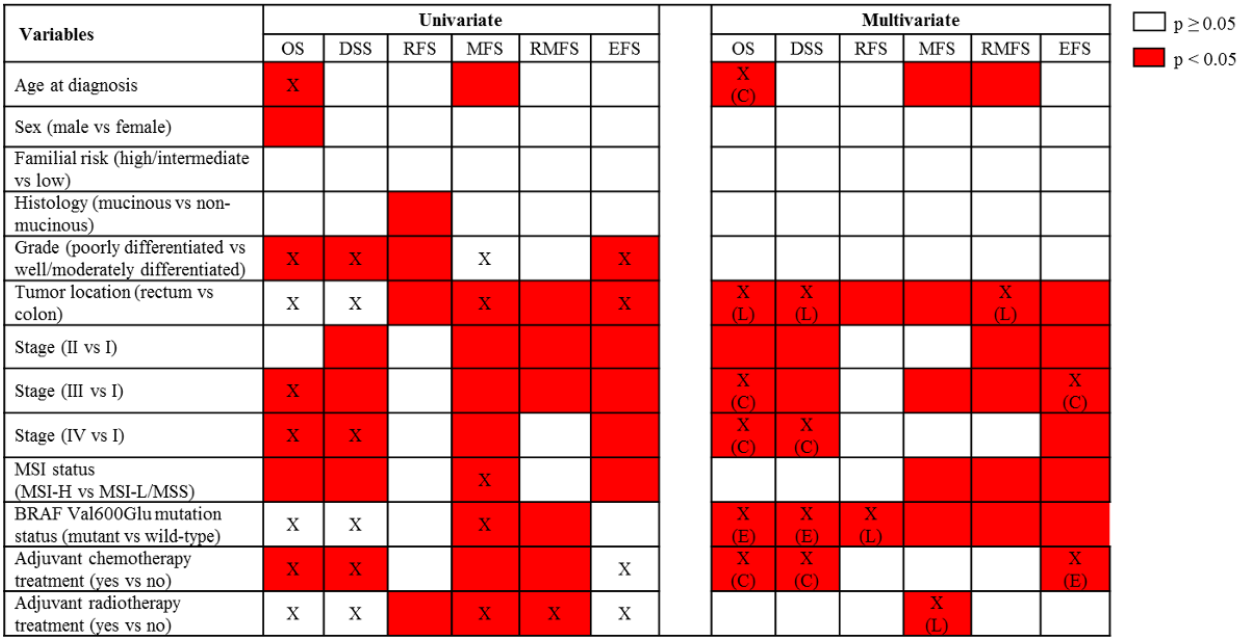


Figure 4.2. Associations between clinico-demographic/molecular markers and the survival outcomes. C, change of effect-size; DSS, disease-specific survival; E, early-effect; EFS, event-free survival; L, late-effect; MFS, metastasis-free survival; MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; OS, overall survival; RFS, recurrence-free survival; RMFS, recurrence/metastasis-free survival. X, variables

that violated the PH assumption in the univariate Cox analyses, or violated the PH assumption and found to have time-varying associations in the multivariable Cox analyses.

4.5.3.2 Multivariable Cox regression models

Multivariable models are shown in **Supplementary Tables 12-17** and main findings are summarized in **Figure 4.2**. Age at diagnosis was associated with overall survival as well as metastasis-related outcomes (**Supplementary Tables 12, 15-16**). The effect size of this variable on the risk of death from any cause became slightly larger after 10.5 years post-diagnosis (**Supplementary Table 12**). In contrast, increasing age at diagnosis was associated with decreased risks of MFS (HR: 0.98; **Supplementary Table 15**) and RMFS (HR: 0.98; **Supplementary Table 16**). Other demographic variables (sex and familial risk) as well as tumor histology and grade were not associated with any of the survival outcomes.

Tumor location and BRAF Val600Glu mutation status were associated with all outcomes. Effects of these two variables either remained constant or varied with time on different disease outcomes (**Figure 4.2**; **Supplementary Tables 12-17**). For example, results of RFS, MFS, and EFS analyses showed that rectal cancer patients compared to colon cancer patients had shorter times to events throughout the follow-up time with no detectable time-varying associations (**Supplementary Tables 14-15, 17**). In the OS analysis, no significant difference between the rectal and colon cancer patients were detected prior to 2 years following diagnosis. However, after this time point, the risk for rectal cancer patients became significantly higher (HR: 1.68; **Supplementary Table 12**). Also, while in the early years RMFS and DSS times did not significantly differ between the rectal and colon cancer patients, after 3 years in RMFS and after

6.5 years in DSS, the event risk significantly increased for the rectal cancer patients (HRs: 3.91 for RMFS and 5.97 for DSS; **Supplementary Tables 13** and **16**). Presence of BRAF Val600Glu mutation was associated with shorter overall and disease-specific survival times within the first 2.5 years post-diagnosis (HR: 2.18 for OS and 3.05 for DSS), but not after that (**Supplementary Tables 12-13**). This mutation was also significantly associated with an increased risk of recurrence after 4 years following diagnosis (HR: 7.10; **Supplementary Table 14**). Last, patients with this tumor mutation had shorter MFS, RMFS, and EFS times without any time-varying associations (**Supplementary Tables 15-17**).

Stage was associated with all outcomes except the risk of recurrence (**Figure 4.2**). Patients with advanced stages had generally increased risks of outcome events, stage IV disease was a strong predictor of death, and stage III disease was a predictor of metastasis (**Supplementary Tables 12-13, 15-17**). For this variable, time-varying associations were found on death-related outcomes (i.e. OS, DSS, and EFS). Specifically, compared to stage I patients, stage III and stage IV patients had a much higher risk of death from any cause within the 1st year following diagnosis than later (**Supplementary Table 12**). Similar to this, for stage IV patients the risk of death from colorectal cancer was much higher during the 1st year post-diagnosis (**Supplementary Table 13**). Additionally, the risk of having at least one disease-related events for stage III patients (EFS) was much higher within the first 1.5 years following diagnosis, which then decreased in magnitude (HRs: 6.02 within the first 1.5 years post-diagnosis versus 2.99 after that; **Supplementary Table 17**). Effects of disease stage on other outcome measures did not change over time.

Tumor MSI phenotype was associated with only metastasis-related outcome measures (MFS, RMFS, and EFS; **Figure 4.2**). MSI-H tumor phenotype had a protective effect

(**Supplementary Tables 15-17**). Unlike other variables, MSI status had no time-varying associations.

Last, the two treatment-related variables, adjuvant chemotherapy and adjuvant radiotherapy treatment statuses, showed different association patterns in this observational cohort. While associations of the adjuvant chemotherapy treatment was detected in death-related outcomes (i.e. OS, DSS, and EFS), adjuvant radiotherapy treatment was only associated with MFS (**Figure 4.2**). These effects were non-proportional during the follow-up (i.e. varied over time). Specifically, within the 1st year following diagnosis, adjuvant chemotherapy had strong and significant protective effects on OS, DSS, and EFS, after which this effect was not detectable in the EFS analysis, but was still significant in the OS and DSS analyses, albeit with a decreased effect size (HR: 0.05 for OS, 0.15 for DSS, and 0.40 for EFS within the first year post-diagnosis, and 0.56 for OS and 0.50 for DSS after this time-point) (**Supplementary Tables 12-13** and **17**). Whereas, compared to patients who did not receive adjuvant radiotherapy, patients who received adjuvant radiotherapy had an increased risk of metastasis (HR: 6.00) after 5.5 years following diagnosis (**Supplementary Table 15**).

4.6 Discussion

In this study, we examined the survival characteristics of a prospective cohort of colorectal cancer patients (n=738) followed up to 19 years and association of a set of baseline variables with outcome measures. This long follow-up time makes it an excellent resource for investigation of prognostic characteristics in both the short- and long-term. Our results show the

survival characteristics in this patient cohort over a long follow-up time; describe the relationships between baseline clinical, demographic, and select tumor molecular markers and a comprehensive set of patient clinical outcomes; present interesting findings regarding variables with time-varying associations; and identify a set of candidate early-effect and late-effect markers that can help distinguish patients who are at increased outcome risks during specific time periods following diagnosis.

4.6.1 Long-term survival characteristics of the patient cohort

Overall, some of our results supported previous literature findings and some others provided new insights. Characteristics of the patient cohort and survival probabilities are shown in **Tables 4.1-4.2** and **Figure 4.1**. As expected, a portion of the patients experienced disease progression (i.e. recurrence/metastasis) and this was strongly linked to death from colorectal cancer. The majority (85%) of the first recurrence and/or metastasis (**Figures 4.1C-E**) and deaths from colorectal cancer (**Figure 4.1B**) were clustered during the first ~4.5-5 and ~6 years, respectively. These findings, similar to other reports, emphasize the initial years after diagnosis as a critical window of time for colorectal cancer patients [79,384,385]. However, in some patients the first recurrence or metastasis happened after the first 5-years (15.8% and 13.3% of the events, respectively). This raises the question of whether the medical surveillance should be extended beyond the most recommended time frame of 5-years for the patients who did not experience disease progression until then. Similar observations and suggestions were made by others [80,81]. On the positive side, our results (**Figure 4.1F**) also showed that when a patient survived the initial 6 years without any disease-related event (recurrence, metastasis, or death from colorectal

cancer), their risk for these disease outcomes became much less afterwards (~95% of the patients who had any of these events had their first events or died within the first 6 years). This suggests that the long-term consequences of colorectal cancer become minimal once a patient survives the first 6 years event-free.

4.6.2 Modeling time-varying associations and previous literature findings in colorectal cancer

In order to examine the relationships between the variables and outcome measures, we applied both the univariate and multivariable analyses. In these analyses, we aimed to explore the variables for their constant as well as potential time-varying associations. We note that while the term "effect" suggests a direct effect of the variable, it should not be taken literally – it rather reflects an association. In our case, variables with constant effects are those that satisfy the proportional hazards assumption of the Cox model and for which the hazard ratio estimations throughout the follow up time remain constant. Variables with time-varying associations, on the other hand, are those that have their effects (i.e. HRs) change over time. This also means that a marker's effect may only be detectable or obvious during a specific time period, or the direction of the marker's effect may change over different time-periods. Intuitively, to identify such variables, data obtained from cohorts followed up for a long time, like the cohort examined in this study, is needed. Previous studies reported that age, sex, grade, stage, tumor location/site, a somatic alteration, and a few genetic polymorphisms had potential time-varying associations in colorectal cancer [177,191–196,395,396]. However, to our knowledge, only a few of these studies identified the time periods using the patient data, which reflect the patterns of effects on survival

times [177,196,395], as we did in this study. Also, in our study we used Cox model with time varying effects assuming piece-wise constant hazard ratios, which provided simple (i.e. one time-point per variable) and potentially clinically meaningful information.

4.6.3 Time-varying associations identified in the univariate analyses and implications for multivariable modeling

In our study, distinct patterns of survival probability for variables with non-proportional effects (Type A and Type B variables) were observable after univariate analyses and assessment of the PH assumption (**Supplementary Figures 4-6**). It should be noted that the differences between the Type A and Type B variables have implications for researchers: characteristics of the Type B variables (i.e. which do not have a significant p-value in the univariate analyses) indicate that such variables may be excluded from multivariable modeling if the researchers select the covariates based on the univariate p-values. Such an exclusion could then lead to omission of important variables (e.g. those with potential time-varying associations) in the final models.

4.6.4 Multivariable models and associations detected with or without time-varying associations

4.6.4.1 Demographic factors and their relation to outcome measures

Multivariable models that considered the time-varying associations yielded a number of interesting findings (**Supplementary Tables 12-17**). Regarding the demographic features, increasing age at diagnosis was associated with a small but significant increased risk of mortality throughout the follow-up time. This risk became slightly larger after the initial 10.5 years (OS; **Supplementary Table 12**). It is not surprising that younger patients had a lower risk of death, as they normally would have fewer comorbidities, lower chances of dying from other causes, and are likely to receive aggressive and intense treatments [398] that may contribute to their longer survival times. The slight increase in the risk of death after a decade can be explained by aging of the patients in the cohort. On the other hand, small but long-term effects were detected for age at diagnosis on metastasis-related outcomes (MFS and RMFS) where decreased age was associated with worse MFS/RMFS times (**Supplementary Tables 15-16)**. It is reported by other studies that younger colorectal cancer patients present with advanced diseases [398,399]. In our cohort 32.8% and 27.7% of the younger patients (age at diagnosis < 65) and older patients, respectively, were diagnosed with a stage III disease - this may explain the increased metastasis risk in the young patients. In contrast to age, another demographic variable, sex, was not associated with any of the survival outcomes examined in this study. The role of patient sex in prognosis is controversial: some studies support that female patients have better prognosis compared to male patients [400–402] while others do not find such a sex-based difference [190,403]. We observed a better survival for female patients in the univariate analysis, but this association was

not retained in the multivariable models. Additionally, consistent with other studies [389,404], familial risk status, while it is a risk factor for development of colorectal cancer [390,405], had no significant relation to any survival outcomes investigated. Therefore, in this cohort age at diagnosis has emerged as the only demographic factor with a predictive role.

4.6.4.2 MSI and disease stage and their relation to outcome measures

In our analysis MSI status was predictive of only metastasis-related outcomes (MFS, RMFS, and EFS; **Supplementary Tables 15-17**) and its effects remained stable during the entire follow-up. MSI-H is a known marker with protective effects on patient survival [387,406], possibly due to its biological effect on metastasis through its association with increased immune cell infiltration [150]. Thus, our results are consistent with these previous findings but additionally emphasize that the MSI status predicts the risk of metastasis even long after the diagnosis. As the most important prognostic marker, stage was a predictor of the majority of the outcome measures investigated (**Figure 4.2**). As expected, increased disease stage was generally associated with increased risk of events, but in some cases the hazard ratios significantly differed before and after a time-point. Interestingly, such effects were detected in death-related outcomes. Specifically, fluctuating HRs were detected for stage III patients in the OS and EFS analyses and for stage IV patients in the OS and DSS analyses. In these cases, the risk of event was much higher for the patients immediately after the diagnosis (i.e. within the 1 - 1.5 years) compared to later. This time-relationship may be attributed to the advanced disease at diagnosis and/or the post-surgical complications that are known to lead to early death [407–409]. We note that in a previous study on OS, similar findings (i.e. non-proportional effects of stage III and stage IV disease) were reported [195].

4.6.4.3 Tumor location and its relation to outcome measures

Two variables were associated with all outcomes examined in this study and tumor location was one of them. Tumor location is one of the most widely examined clinical variables in colorectal cancer and is used in the clinic for prognostic estimations as well as surveillance and treatment-related decisions. In our study, rectal tumors compared to colon tumors were associated with worse RFS, MFS, and EFS times throughout the follow-up with no time-varying association (**Supplementary Tables 14-15, 17**). It is known that the rectal cancer patients have a higher risk of recurrence and metastasis [410,411], which is also shown by our results. However, our results additionally showed that the rectal tumors had sustained these constant/continuous effects over a long time after the diagnosis. In contrast, in the OS, DSS, and RMFS analyses, we observed time-varying associations of tumor location. The DSS and RMFS model data were particularly interesting. In both models, rectal cancer patients tended to have worse outcomes compared to colon cancer patients, but this difference reached significance only after certain time-points. In the RMFS model, the risk for increased recurrence/metastasis became significantly higher for the rectal cancer patients only after the initial 3 years. Since recurrence and metastasis indicate disease progression, RMFS data may be particularly relevant for clinical surveillance purposes and may suggest that the rectal cancer patients who survived the first 3 years without disease progression may need to be carefully surveilled after this time period. Additionally, a similar and a later effect was observed in the DSS model, where the risk of death from colorectal cancer significantly increased for the rectal cancer patients after 6.5 years. The non-proportional effect of tumor location on DSS has been observed by others as well [190]. In our study, the increased risk of disease progression for rectal cancer patients after 3 years

(**Supplementary Table 16)** may explain their increased risk of disease-specific death after 6.5 years (**Supplementary Table 13**).

4.6.4.4 BRAF Val600Glu mutation and its relation to outcome measures

Like tumor location, BRAF Val600Glu mutation status was associated with all outcome measures (**Supplementary Tables 12-17**). This mutation is one of the most studied tumor mutations in colorectal cancer as well as other cancer sites, such as ovarian cancer [412], thyroid cancer [413], lung cancer [414], and melanoma [388]. In our study, patients with this tumor mutation had increased risk of two metastasis-related outcomes throughout the follow up (the highest risk being associated with metastasis-free survival; HR: 3.46). Such a relationship between mutant BRAF and metastasis was previously reported in other cohorts [415,416]. This mutation was also associated with shorter time to recurrence after 4 years. It is not immediately clear how this mutation may influence the recurrence risk in colorectal cancer, but the association of this mutation with tumor recurrence has been reported in papillary thyroid cancer as well [413]. In addition to these, previously BRAF Val600Glu mutation has been associated with the increased risk of mortality in colorectal cancer [156,159,392,416,417]. In our study, in two death-related outcomes (OS and DSS), this mutation emerged as a predictor of death early after diagnosis (within the first 2.5 years). Interestingly, this group of patients also tended to have better DSS times if they survived the initial 2.5 years following diagnosis, but this did not reach significance levels (HR: 0.14, p=0.0505; **Supplementary Table 13**). BRAF Val600Glu mutation status is the only variable identified in this study that was both an early-event (OS and DSS) and late-event (RFS) marker. The reason why this mutation has such effects remains unknown and warrants more

investigations. Overall, the wide-spectrum of associations detected for this mutation in this study further strengthen this gene's importance in colorectal cancer.

4.6.4.7 Adjuvant chemotherapy and radiotherapy treatment status and their relation to outcome measures

Adjuvant therapy is given based on the clinical and disease characteristics to help control the disease (e.g. to reduce/eliminate the recurrence and/or metastasis risk) and to improve the survival outcomes of patients. In our patient cohort, patients who received adjuvant chemotherapy had better survival outcomes (OS, DSS, and EFS) than those patients who did not receive it. These effects were especially stronger within the first year following diagnosis (OS, DSS, and EFS models; **Supplementary Tables 12-13, 17**). The changing-effects (from strong to weaker protective effects) may reflect the slightly diminishing effects of therapy after the treatment duration, which is usually no more than a year [418]. Time-varying associations for chemotherapy treatment were detected in other cancers as well, such as breast cancer [419–421]. On the other hand, adjuvant radiotherapy status was associated with only MFS (**Supplementary Table 15**). Initially MFS times did not differ significantly for the patients who did or did not receive this treatment. However, after 5.5 years following diagnosis, those patients who received radiotherapy had increased risk of developing their first metastases compared to patients who did not receive this treatment. The exact mechanisms through which adjuvant radiotherapy can have a late effect on MFS of patients is not clear, but it is known that in some cases radiation treatment increases the risk of metastasis [422–424]. As these authors discussed [422–424], a variety of potential mechanisms can explain this effect, such as the appearance or development of radiation

resistant tumor cells, changes in tumor microenvironment or immune system response over time, or suppression of the tumor progression by radiation treatment that initially delays the tumor metastasis. These previous and our findings emphasize the need for new research revenues and potentially prolonged surveillance for late-onset metastatic lesions in colorectal cancer patients who are treated with adjuvant radiotherapy.

4.6.5 Strengths and limitations

Limitations of this study include the missing information on the cause of death for a portion of the patients; assuming that the non-colorectal cancer related deaths were independent of colorectal cancer; having a small number of recurrences in the dataset, which may have limited the study power in analysis of recurrence-related outcomes; and examining select clinico-demographic and tumor molecular markers, which leaves it to future studies to examine the potential effects of other markers. Additionally, characteristics of the patients who are included in this study may differ from the patients who were diagnosed during the recruitment phase, but declined to consent and participate in NFCCR. This may affect the generalizability of the findings. However, it should also be noted that in some cases the consent to access the medical records and tissue specimen was obtained from the close relatives/proxies of the patients who had died. Thus, bias that may be introduced by exclusion of advanced stage patients is expected to be lower in our study compared to many other studies [329]. This study also has a number of unique advantages: the cohort examined in this study is one of the longest followed-up cohorts that allowed the systematic examination of long-term survival characteristics in colorectal cancer; this is a prospective cohort study that reduces information bias compared to retrospective

cohort studies [425]; a comprehensive set of outcome measures were examined, which provided detailed information on survival patterns and relationships; and finally, the PH assumption in Cox regression models was checked and effects of variables were properly assessed - this not only increased the reliability of the effect-estimations, but also allowed us to identify promising early and late effect markers.

4.7 Conclusions

In conclusion, this study describes the long-term survival characteristics of a prospective cohort of colorectal cancer patients and the detailed relationships between baseline variables and patient outcomes over a long time. Overall, our results increase the depth of information on patient outcomes and the markers of short-term and long-term risks, and provide new insights that may assist future research and clinical care strategies in colorectal cancer.

4.8 Acknowledgements

The authors thank the patients for making this study possible. We also wish to thank all the personnel and investigators who contributed to the patient recruitment and data collection at NFCCR as well as the staff at NLCHI and Provincial Tumor Registry-NL/Dr. H. Bliss Murphy

Cancer Centre for their assistance with a portion of the patient outcome data. SS is a senior investigator of the Beatrice Hunter Cancer Research Institute (BHCRI).

4.9 Ethics approval and consent to participate

Ethics approval (reference numbers 09.106 and 15.294) was provided by the Health Research Ethics Board (HREB) of Newfoundland and Labrador prior to start of this study. Since this is a secondary-use-of-data study, HREB has waived the need for patient consent.

CHAPTER 5: A comprehensive analysis of SNPs and CNVs identifies novel markers associated with disease outcomes in colorectal cancer

A version of this manuscript has been published in Molecular Oncology, 2021, doi: 10.1002/1878-0261.13067. Note that supplementary information that was published with the manuscript is presented in Appendix E.

Yajun Yu[1], Salem Werdyani[1], Megan Carey[1], Patrick Parfrey[2], Yildiz E. Yilmaz[1,2,3], Sevtap Savas[1,4]

[1] Discipline of Genetics (As of Sep 2020, the Discipline of Genetics has become a part of the Division of Biomedical Sciences), Faculty of Medicine, Memorial University, St. John's, NL, Canada.

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

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

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

5.1 Co-authorship statement

Yajun Yu helped design the statistical approach, performed the imputations, conducted all statistical and bioinformatics analyses, interpreted the results, and drafted the manuscript.

Salem Werdyani generated the CNV and INDEL data analyzed in this study.

Megan Carey helped collect the outcome data.

Patrick Parfrey led the NFCCR.

Yildiz E. Yilmaz conceptualized the study, led the statistical design.

Sevtap Savas conceptualized, led, and helped design the study, helped collect patient-related data, helped draft and revised the manuscript, and submitted the manuscript.

5.2 Abstract

We aimed to examine the associations of a genome-wide set of single-nucleotide polymorphisms (SNPs) and 254 copy number variations (CNVs) and/or insertion/deletions (INDELs) with clinical outcomes in colorectal cancer patients (n=505). We also aimed to investigate whether their associations changed (e.g. appeared, diminished) over time. Multivariable Cox proportional hazards and piece-wise Cox regression models were used to examine the associations. The Cancer Genome Atlas (TCGA) datasets were used for replication purposes and to examine the gene expression differences between tumor and non-tumor tissue samples. A common SNP (*WBP11*-rs7314075) was associated with disease-specific survival with p-value of 3.2×10^{-8}. Association of this region with disease-specific survival was also detected in the TCGA patient cohort. Two expression quantitative trait loci (eQTLs) were identified in this locus that were implicated in the regulation of *ERP27* expression. Interestingly, expression levels of *ERP27* and *WBP11* were significantly different between colorectal tumors and non-tumor tissues. Three SNPs predicted the risk of recurrent disease only after 5-years post-diagnosis. Overall, our study identified novel variants, one of which also showed an association in the TCGA dataset, but no CNVs/INDELs, that associated with outcomes in colorectal cancer. Three SNPs were candidate predictors of long-term recurrence/metastasis risk.

5.3 Introduction

A significant portion of colorectal cancer patients die of this disease, and develop local recurrences and metastases over time [79,426]. Knowledge on the baseline predictors of clinical outcomes is essential for effective disease management. The disease stage is the most well-known prognostic marker in colorectal cancer [137,138]. Other factors, including tumor location, microsatellite instability status, and treatment have also been associated with patient outcomes [387,427,428]. However, patients who are categorized in the same prognostic group may experience different outcomes, indicating the need for additional prognostic markers to distinguish between patients with different outcome risk. Given that genetics plays a role in many human phenotypes, it is intuitive to hypothesize that genetic variants can be prognostic markers in colorectal cancer.

A number of studies have examined the associations of genetic variations, such as SNPs, with clinical outcomes in colorectal cancer. While these studies focused mostly on candidate variant, gene, or pathway analyses [167–174,429–433], a small number of genome-wide association studies (GWASs) were also performed [176–179,181,434]. These GWASs focused on often diverse outcome measures, identified a limited set of variants and potential genes, and their results largely remain to be confirmed by further studies. SNPs are the most common genetic variables, however, human genome also contains copy number variants (CNVs; ≥1 kb) and insertion/deletion variants (INDELs; <1 kb). While analysis of copy number alterations in tumor genomes are widely performed, there are not many studies that have checked the potential associations of germline CNVs/INDELs with survival outcomes in colorectal cancer [292,293,395,396]. As a result, similar to SNP

studies, only a handful of genes and CNVs/INDELs have been identified as candidate prognostic markers in colorectal cancer.

Survival studies can identify prognostic markers that can predict the hazard over the follow-up periods [199,200,218]. Normally, such markers can distinguish between patients with different outcome risk regardless of time. In rare cases, however, it has been shown that some markers have different levels or types of associations during different time-periods of the follow-up (i.e. time-varying associations). Such markers, therefore, can help distinguish between patients with high and low outcome risk during certain time-periods. For example, in our previous colorectal cancer study, prognostic associations became stronger, weaker, appeared, or diminished over time for a set of baseline clinical variables [189]. Similarly, we and others identified two somatic alterations [189,196] and three genetic polymorphisms [177,395] that were associated with early or late risk of disease outcomes in colorectal cancer. Knowledge on such markers is surprisingly limited. This may be because that many cohorts do not have long follow-up times that are essential for identifying whether a variable has constant or time-varying associations with outcomes.

This literature information indicates that further studies on genome-wide sets of SNPs, CNVs/INDELs, and colorectal cancer outcomes are necessary to improve the current level of knowledge. In addition, there is a need for studies that investigate time-varying associations, as this type of analysis provides unique insight into prognosis. In this study, we examined large sets of common genetic variants (~ 4.7 million SNPs and 254 CNVs/INDELs) and their associations with disease-specific survival and recurrence/metastasis-free survival in a colorectal cancer patient cohort (n=505 and 495, respectively) followed up to 19 years. Our objectives were to: 1) investigate the

associations of genetic variants with the outcomes, 2) examine whether any of the variants had time-varying associations, and 3) further explore our findings using The Cancer Genome Atlas (TCGA) datasets for replication purposes and gene expression analyses.

5.4 Methods

5.4.1 Ethics approval

This study complied with the Declaration of Helsinki and was approved by the Human Research Ethics Board (HREB) of Newfoundland and Labrador (reference numbers: 2009.106; 2015.294; 2016.252). As this is a research study with a secondary use of data, HREB waived the consent requirement.

5.4.2 Patient cohort, and clinical and genetic data

Patients in the Newfoundland Colorectal Cancer Registry (NFCCR) cohort were diagnosed between 1999 and 2003 and followed up to 19 years **(Appendix E)** [189,262,328,329]. DNA samples extracted from white blood cells were available for 539 patients at the time of genotyping. Out of 539, patients who passed the sample quality control measures, satisfied the inclusion criteria [176], and had the genetic data available

(SNP or CNV/INDEL genotype data) were included in the analyses. All patients included were Caucasians and unrelated to each other [176].

Genetic data examined in this study includes two datasets [176,395]. The SNP dataset, which is available for 505 patients (**Table 5.1**), includes 4,711,309 SNPs that qualified for analysis (genotyped SNPs=607,365; imputed SNPs=4,103,944). Genetic imputation was done using SHAPEIT (v2.r837) [435] and IMPUTE2 (v2.3.2) [436], using the 1000 Genomes Phase 3 data [235] as the reference panel data. The initial SNP genotype data, inclusion/exclusion and quality control (QC) metrics, and imputation procedures are explained in detail in **Appendix E**. Quality control measures were applied to variants: info scores of imputed SNPs >0.7, maximum probability of the imputed genotypes >0.9, and for all SNPs in the dataset, Minor Allele Frequency (MAF) ≥10%, missing genotype data rates (for SNPs and individuals) ≤5%, and Hardy-Weinberg Equilibrium (HWE) p-value $>1\times10^{-08}$. All imputed SNPs included in the statistical analyses had an info score >0.8. For simplicity we refer to the genetic variants in this dataset as "SNPs", even though the genotyping platform and imputation results contain other variant types, such as INDELs.

Table 5.1. Baseline characteristics of the SNP and CNV/INDEL analysis cohorts.

	SNP analysis cohort (n = 505)		CNV/INDEL analysis cohort (n = 495 *)	
Variable	Number	%	Number	%
Age at diagnosis				
Median (range)	61.43 (20.70-75.01)	-	61.40 (20.70-75.01)	-
Sex				
Male	307	60.79	301	60.81

Female	198	39.21	194	39.19
Tumor location				
Colon	334	66.14	328	66.26
Rectum	171	33.86	167	33.74
Stage				
I	93	18.42	89	17.98
II	196	38.81	193	38.99
III	166	32.87	164	33.13
IV	50	9.90	49	9.90
Histology				
Non-mucinous	448	88.71	438	88.48
Mucinous	57	11.29	57	11.52
Grade				
Well/moderately differentiated	464	91.88	457	92.32
Poorly differentiated	37	7.33	34	6.87
Unknown	4	0.79	4	0.81
MSI status				
MSI-L/MSS	431	85.35	421	85.05
MSI-H	53	10.50	53	10.71
Unknown	21	4.16	21	4.24
***BRAF* Val600Glu mutation**				
Wild-type	411	81.39	402	81.21
Mutant	47	9.31	47	9.49
Unknown	47	9.31	46	9.29
Adjuvant chemotherapy treatment				
No	224	44.36	217	43.84
Yes	277	54.85	274	55.35
Unknown	4	0.79	4	0.81
Adjuvant radiotherapy treatment				
No	364	72.08	355	71.72
Yes	124	24.55	123	24.85
Unknown	17	3.37	17	3.43
Follow-up time				
Median (range)	13.79 (0.38-19.00)	-	13.80 (0.38-19.00)	-
DSS status				
Death from other causes or alive	332	65.74	323	65.25
Death from colorectal cancer	99	19.60	99	20.00
Unknown	74	14.65	73	14.75
Death from other causes or alive (within 5 years)	407	80.59	398	80.40

Death from colorectal cancer (within 5 years)	62	12.28	62	12.53
Unknown (within 5 years)	36	7.13	35	7.10
[#] **RMFS status**				
Recurrence or metastasis (−)	331	72.75	322	72.20
Recurrence or metastasis (+)	124	27.25	124	27.80
Recurrence or metastasis (−) (within 5 years)	348	76.48	339	76.01
Recurrence or metastasis (+) (within 5 years)	105	23.08	105	23.54
[$] Unknown (within 5 years)	2	0.44	2	0.45

CNV, copy number variation; DSS, disease-specific survival; INDEL, insertion/deletion; MSI, microsatellite instability; MSI-H, microsatellite instability-high; MSI-L, microsatellite instability-low; MSS, microsatellite stable; RMFS, recurrence/metastasis-free survival; SNP, single nucleotide polymorphism.

*, Note that all 495 patients in the CNV/INDEL analysis cohort are also in the SNP analysis cohort with 505 patients. [#], Stage I-III patients only, total n = 455 in the SNP analysis cohort and total n = 446 in the CNV/INDEL analysis cohort. [$], 'Unknowns' appear because two patients had unknown survival time. Although they experienced recurrences/metastases, we do not know whether they had these events within the first 5 years postdiagnosis or after that.

In addition to the outcome measures examined, the SNP dataset largely differs from the dataset that we used in a previous genome-wide association study [176] (due to the imputation that allowed us to obtain genotypes of additional variants and the use of longer follow-up data in this study).

The second genetic dataset consists of a set of CNVs/INDELs (**Supplementary Table 18**) [395]. The CNV/INDEL dataset (n=3,486) was previously obtained by our team [395] using a computational pipeline that included PennCNV [353] and QuantiSNP [352] software (**Appendix E**), and was available for 495 patients (**Table 5.1**). These 495 patients were also included in the SNP dataset cohort described above. 254 CNVs/INDELs (**Supplementary Table 18**) that passed filtering based on having copy number state of 0 (i.e. homozygous deletion) in 10-90% in the patient cohort were

analyzed. We had previously examined the associations of 106 of these CNVs/INDELs in the patient cohort with a different outcome measure defined based on a shorter follow-up data [395].

5.4.3 Statistical analyses

5.4.3.1 Correlation among the variables

LD r^2 values were calculated for genetic variants using PLINK v1.07 [437]. Pair-wise Pearson correlation coefficient (r) values were calculated for baseline variables (**Supplementary Table 19**), which suggested that no collinearity ($r < 0.8$) existed among these variables.

5.4.3.2 Outcome measures

The outcome measures are disease-specific survival (DSS) and recurrence/metastasis-free survival (RMFS). Endpoint events in these outcome measures are death from colorectal cancer, and local recurrence or distant metastasis, respectively. DSS and RMFS times are calculated as the times from the date of diagnosis till the date of the occurrence/diagnosis of these events, or the date of last alive contact. DSS was examined for stage I-IV patients and RMFS was analyzed for stage I-III patients only (**Table 5.1**).

5.4.3.3 Survival analysis

Univariate Cox models were fitted for 4,711,309 SNPs for both outcome measures separately. The proportional hazards (PH) assumption was tested under the univariate Cox models using the cox.zph function of the survival package [205] in R [206]. SNPs that satisfied the PH assumption (p-value of the PH assumption test ≥0.05) were then checked for their Cox regression p-values. Those with p-values $<5\times10^{-06}$ were retained for multivariable analysis (**Supplementary Figures 7-8**). On the other hand, SNPs that violated the PH assumption (i.e. variants with possible time-varying associations) were re-fitted in univariate piece-wise/change-point Cox PH regression models [210,383] with a time point of 5 years as the cut-off time point. Five-years was chosen as the time-point to help practically fit a large number of SNPs that violate the PH assumption while also providing a clinically meaningful time-point. PH assumption was then checked for these SNPs before and after the 5 years cut-off time point. Those that satisfied the PH assumption at both time intervals and had Cox regression p-values $<5\times10^{-06}$ before and/or after 5 years post-diagnosis were selected for multivariable analysis (**Supplementary Table 20**). Select Manhattan, regional, and QQ plots are depicted in **Supplementary Figures 7-12**. The genomic regions/loci with independent association signals are defined as ±500 kb of the identified variants with the smallest p values (i.e. index variant), while also considering the LD information (other identified variants in these regions should have $r^2 \geq 0.8$ with the index variants).

Covariates used to adjust the associations of SNPs in multivariable models were identified through the process of baseline model construction. In short, baseline models were constructed using the backward selection method (considering the clinical variables shown in **Table 5.1**) as described in Yu et al. [189], followed by force entering the adjuvant chemotherapy and adjuvant radiotherapy statuses. When the PH assumption for a clinical variable was violated, proper cut-off time points were considered. Further details of this process is shown in **Appendix E**. In the end, tumor location (with a cut-off time point of 6 years), disease stage, microsatellite instability (MSI) status, adjuvant chemotherapy, and adjuvant radiotherapy (with a cut-off time point of 7 years) were remained in the final baseline model for DSS. For RMFS analysis, tumor location (with a cut-off time point of 3 years), disease stage, and adjuvant chemotherapy and radiotherapy treatments were included in the final baseline model.

These baseline variables were then used as covariates in multivariable analysis adjusting the association of variants with survival outcomes. Principal component analysis in the patient cohort did not indicate population stratification (the top principal component accounted for merely 0.3% of the total variance), hence, principal components obtained from the genetic data were not included as covariates. The final multivariable Cox models are the ones with the PH assumption satisfied for all variables (**Appendix E**). Hazard ratios (HRs) and 95% confidence intervals (CIs) were obtained from the multivariable Cox models.

SNPs in this study were examined under additive, dominant, and recessive genetic models. We included recessive model in order not to miss potential associations, however, results should be taken with caution because of the rarity of the homozygous

genotypes. Variants with Cox regression p-values $<5\times10^{-08}$ (either during the entire follow-up [i.e. with no time-varying associations], or before and/or after 5 years post-diagnosis [i.e. with time-varying associations]) were considered to be the variants that were significantly associated with the survival outcome.

Statistical analysis of the CNV/INDEL dataset followed the same analysis procedure as the SNP dataset. During the statistical analyses, patients with homozygous deletions were compared with the patients with other copy number states (i.e. ≥1 copy of the variant).

The empirical power (based on 10,000 simulation replicates) was calculated using the survSNP package [438] in R [206]. This study has at least 80% power to detect effect sizes of 3.2, 3.6, and 18.4 (in DSS analysis) and 3.0, 3.4, and 16.8 (in RMFS analysis) under the additive, dominant, and recessive models, respectively, for variants with a MAF of 10%. Generally, increased power is expected as MAF increases. We expect the same power for the first interval (i.e. the first 5 years post-diagnosis), but less power for the second interval, as the number of events is less in that time period.

Statistical analyses were performed using R ver. 3.5.0 [206] unless otherwise specified. Kaplan-Meier curves, Manhattan and QQ plots were generated using the survival [205] and qqman [439] packages in R [206], respectively. Regional plots were created using software LocusZoom [440].

5.4.4 Validating associations in the TCGA cohort

White (excluding Hispanics/Latinos) colorectal cancer patients with primary tumors were selected. Clinical and outcome data were downloaded from the Genomic Data Commons (GDC) data portal [441] (https://portal.gdc.cancer.gov/; nationwidechildrens.org_clinical_patient_coad.txt, nationwidechildrens.org_clinical_patient_read.txt, nationwidechildrens.org_auxiliary_coad.txt, nationwidechildrens.org_auxiliary_read.txt) (on Dec 13 - 14, 2020) and a study published in 2018 [442], respectively. Germline genetic data of patients (obtained from blood) were obtained from birdseed files in the GDC Legacy Archive [441] (on Nov 16, 2020). High confidence genotype-calls (birdseed confidence value < 0.1) of SNPs were extracted, and those genotypes with low confidence calls were set as "missing". As a result, clinical and genetic data were available for 266 patients. Among the 266 patients, four were removed because they either had a high heterozygosity rate, or were possible relatives, population outliers, or non-European (**Appendix E**). The final TCGA cohort consisted of 262 unrelated colorectal cancer patients (**Supplementary Table 21**).

Genotypes for the SNP identified in the patient cohort (*WBP11*-rs7314075) were not available in this cohort, but genotype data were available for six SNPs (rs11056174, rs2041909, rs2041908, rs6488711, rs2241221, rs11835363) that are in high-LD with it ($r^2 > 0.8$ based on the European data (EUR) in Haploreg 4.1 database [443]). Genotype data of these SNPs were used to examine their associations with DSS in multivariable Cox models with disease stage, tumor location, MSI status, and the top principal component

as the covariates (**Appendix E**). In all Cox models, PH assumption was checked and satisfied for both the clinical and genetic variables.

Among the 12 SNPs in three loci identified under the recessive genetic model in DSS analysis and their 28 high-LD SNPs, one identified SNP rs12757197 (also named as kgp2690683 in the NFCCR cohort) and three high-LD SNPs (rs358347, rs357167, rs165269) were included in the TCGA dataset. However, these SNPs either had no genotypes with double minor alleles (rs12757197), or had no reliable effect estimations (rs358347, rs357167, rs165269 had "infinity" appeared in their upper limit of the 95% confidence interval) in the analysis using the TCGA data.

5.4.5 Examining the associations of CMS with SNPs in high LD with rs7314075 and *WBP11* expression levels in the TCGA dataset

As per the recommendation of one of the reviewers, we also checked the associations between the genotypes of the SNPs in high-LD with *WBP11*-rs7314075 as well as the *WBP11* tumor gene expression levels with tumor consensus molecular subtypes (CMS) in the TCGA dataset. *WBP11* expression data were downloaded from the UCSC Xena [444] and tumor CMS information was obtained from a study published in 2015 [445]. Fisher's exact test was utilized for testing the association of SNP genotypes with the CMS classifications and Kruskal-Wallis test was used to examine the associations of *WBP11* gene expression levels and CMS classifications (ANOVA was not used because the normality assumption was violated). When a significant association was

detected by the Kruskal-Wallis test, further pair-wise comparison was performed using Dunn's test to see which two CMS groups have different *WBP11* expression levels.

5.4.6 Bioinformatics analyses

The functional consequences of the SNPs identified (and SNPs that are in high-LD with them according to the Haploreg database v4.1 [443], based on the European population) were checked in the RegulomeDB database (v2.0) [446] and GTEx (data release v8) [447] (GTEx had data for colon, but not rectum tissues). Expression levels of genes in tumors and adjacent normal tissues (noted as "solid tissue normal" in TCGA) were explored in UCSC Xena [444] using the colorectal tissue data from TCGA [448]. The gnomAD database [236] was used to search for SNP frequencies in different populations. Official gene names and basic definitions of gene functions were retrieved from Gene Entrez [449].

5.5 Results

5.5.1 Associations between SNPs and survival outcomes

In this study we examined 505 and 495 Caucasian patients from Newfoundland, Canada, in the SNP and CNV analysis parts, respectively. Patients were followed up to 19 years. During this period, 99 patients had died from colorectal cancer and 124 patients had experienced recurrence and/or metastasis (**Table 5.1**).

Associations ($p < 5 \times 10^{-08}$) that are detected for disease-specific survival (DSS) and recurrence/metastasis-free survival (RMFS) in multivariable analyses are shown in **Table 5.2** and **Supplementary Tables 22** and **23**.

Table 5.2. rs7314075 that is significantly associated with disease-specific survival (DSS) in multivariable analysis under the *dominant* and *additive* genetic models.

Chr	Pos	Minor/ major allele	MAF	Variant type	Info score	Genetic model	#HR (95% CI)	P-value	P-value of the PH assumption test	*Located region
12	14945417	A/G	0.14	Imputed	0.964	**Dominant**	3.36 (2.18, 5.16)	**3.27×10^{-08}**	0.96	Intron of *WBP11*
						Additive	2.65 (1.88, 3.75)	**3.24×10^{-08}**	0.63	

Chr, chromosome; CI, confidence interval; HR, hazard ratio; MAF, minor allele frequency; PH, proportional hazard; Pos, position. #, Hazard ratio was estimated under the dominant genetic model for [AA+AB] vs BB and under the additive genetic model for AA vs AB vs BB, where A is the minor allele and B is the major allele. * Gene annotation is obtained from the UCSC database [450] ("UCSC genes" from the UCSC browser [GRCh37/hg19]). Models are adjusted for MSI status, disease stage, tumor location (6 years as the cut-off time point), adjuvant chemotherapy and radiotherapy statuses (7 years as the cut-off time point for adjuvant radiotherapy).

5.5.1.1 Associations with constant HRs

After adjustment for clinical covariates, one common SNP that locates in an intron of *WBP11* (rs7314075) was significantly associated with the risk of death from colorectal cancer under both the dominant (HR=3.36; p-value=3.27×10^{-08}) and additive (HR=2.65; p-value=3.24×10^{-08}) genetic models (**Table 5.2**). Under the dominant genetic model (**Figure 5.1**), patients with AA or AG genotype had more than three times the risk of death from colorectal cancer compared to patients with GG genotype. Under the additive genetic model, in line with the results of the dominant genetic model, risk of death from colorectal cancer increased more than 1.5 folds as per A allele (i.e. the minor allele). With regard to SNPs examined under the dominant and additive models in the RMFS analysis, none of them reached significant p-values in the multivariable analysis. Top SNPs with suggestive associations for these genetic models are shown in **Supplementary Table 24**.

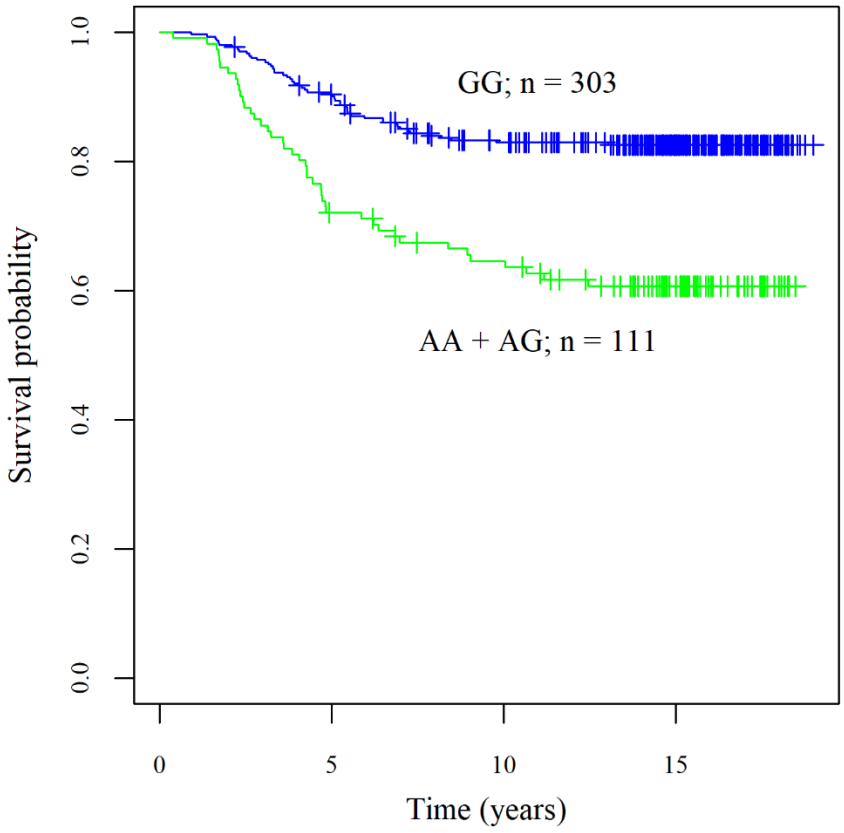


Figure 5.1. Kaplan Meier curves of rs7314075 in the disease-specific survival (DSS) analysis under the dominant genetic model. The p-value of the log-rank test is 2×10^{-06}.

Under the recessive genetic model, associations were detected in multivariable analyses for 13 genomic regions (a total of 12 SNPs from three genomic loci in DSS and 56 SNPs from 10 loci in RMFS analyses) that passed the genome-wide significance level of 5×10^{-08} (p-values 10^{-08}-10^{-12}) (**Supplementary Tables 22** and **23**). Some of these variants were located in genes (**Supplementary Tables 22** and **23**). Since many of these associations included small numbers of minor allele homozygous genotypes, these results should be approached with caution.

5.5.1.2 Time-varying associations

Interestingly, three variants from two different genomic loci (chromosomes 2 and 12: rs200143895, rs11064732, rs817090) had time-varying associations with RMFS under the recessive model after adjustment for clinical covariates. These variants were associated with the risk of recurrence/metastasis only after 5 years post-diagnosis (**Supplementary Table 23**).

No SNPs with time-varying associations were detected in other models examined in multivariable analysis.

5.5.2 Examining the association of *WBP11*-rs7314075 in the TCGA cohort

WBP11-rs7304075 itself was not included in the TCGA genetic data, but there were six SNPs (**Table 5.3**) that were in high linkage disequilibrium (LD) ($r^2 > 0.8$) with it in this dataset. These SNPs were analyzed for their associations with DSS in the TCGA colorectal cancer cohort. Four SNPs (rs11056174, rs2041909, rs6488711, and rs2241221) were significantly associated with the risk of death from colorectal cancer under both the dominant and additive genetic models (adjusted for tumor location, disease stage, MSI status, and the top principal component) (**Table 5.3**). Consistent with the results obtained in our patient cohort (**Table 5.2**), genotypes containing the minor alleles of these SNPs were associated with an increased risk of outcome in the TCGA patient cohort (HRs =

2.93 - 3.00 under the dominant genetic model; HR = 2.32 - 2.39 under the additive model) (**Table 5.3**).

Table 5.3. Associations between SNPs in high-LD with rs7314075 and disease-specific survival (DSS) in multivariable analysis in the TCGA dataset under the *dominant* and *additive* genetic models.

Genetic model	SNP	Chr	Pos	Minor/major allele	MAF	#HR (95% CI)	P-value	P-value of the PH assumption test
Dominant	rs11056174	12	14909977	T/C	0.14	2.94 (1.20, 7.20)	**0.018**	0.56
	rs2041909	12	14915409	C/T	0.14	3.00 (1.23, 7.32)	**0.016**	0.58
	rs2041908	12	14916150	G/A	0.14	2.32 (0.96, 5.65)	0.063	0.73
	rs6488711	12	14933216	T/C	0.14	2.93 (1.20, 7.17)	**0.018**	0.56
	rs2241221	12	14959391	C/T	0.16	2.97 (1.23, 7.16)	**0.015**	0.47
	rs11835363	12	14982700	C/T	0.16	2.42 (1.00, 5.88)	0.050	0.23
Additive	rs11056174	12	14909977	T/C	0.14	2.35 (1.05, 5.29)	**0.038**	0.81
	rs2041909	12	14915409	C/T	0.14	2.38 (1.06, 5.32)	**0.035**	0.85
	rs2041908	12	14916150	G/A	0.14	1.96 (0.87, 4.44)	0.106	0.92
	rs6488711	12	14933216	T/C	0.14	2.32 (1.03, 5.20)	**0.041**	0.79
	rs2241221	12	14959391	C/T	0.16	2.39 (1.08, 5.31)	**0.032**	0.72
	rs11835363	12	14982700	C/T	0.16	2.01 (0.90, 4.50)	0.091	0.39

Chr, chromosome; CI, confidence interval; HR, hazard ratio; MAF, minor allele frequency; PH, proportional hazard; Pos, position. [#], Hazard ratio was estimated under the dominant genetic model for [AA+AB] vs BB and under the additive genetic model for AA vs AB vs BB, where A is the minor allele and B is the major allele. Models are adjusted for MSI status, disease stage, tumor location, and the top principal component.

5.5.3 Functional roles of SNPs

We explored the potential functional features of *WBP11*-rs7314075 and its highly linked SNPs. According to Haplogreg [443], there were 38 SNPs that were highly-linked with the *WBP11*-rs7314075. Two of these highly linked SNPs (rs2241221 and rs11056174) were cis-eQTLs (i.e. located within ±1 Mb region of the transcription start sites of the associated genes) according to RegulomeDB [446] (**Table 5.4**). These SNPs were associated with the expression level of *ERP27* in monocytes. Comparison of gene-expression levels using the TCGA data showed that the expression levels of *ERP27* and *WBP11* were higher in the colon and rectal tumors than in adjacent normal tissues (the "solid tissue normal" in TCGA data) (**Figure 5.2** and **Supplementary Figure 13**).

Table 5.4. Variants that are in high LD with *WBP11*-rs7314075 that are eQTLs.

Outcome - genetic model	rs ID	*eQTL associated gene (tissue) - RegulomeDB	*eQTL associated gene (tissue) - GTEx
DSS-dominant/ additive	rs2241221	*FLJ32115/ERP27* (monocyte)	-
DSS-dominant/ additive	rs11056174	*FLJ32115/ERP27* (monocyte)	-

DSS, disease-specific survival; eQTL, expression quantitative trait locus. *, Variants that are in high-LD with *WBP11*-rs7314075 (retrieved from Haploreg [443]) were explored in RegulomeDB [446] and GTEx [447]. Note that GTEx data were for colon tissue, as it has no data for rectal tissue. The eQTLs are all cis-eQTLs that locate within ±1 Mb of the transcription start sites of the genes shown in the Table.

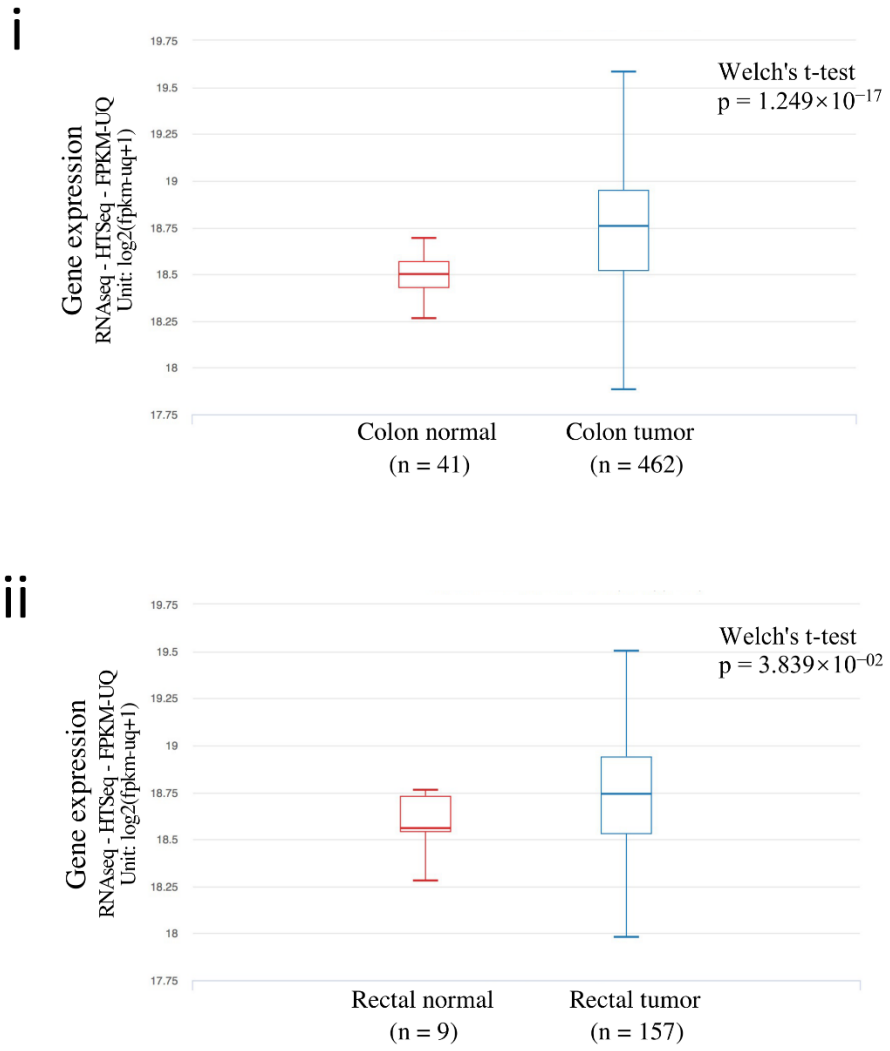


Figure 5.2. Expression level of *WBP11* in colorectal tumors and normal tissues. Analysis was done in UCSC Xena [444] using the GDC TCGA COAD and READ data. In both datasets, primary tumors and adjacent normal tissues (noted as "solid tissue normal" in TCGA data) were selected (recurrent and metastatic tumors were excluded). Then, only tumors and normal tissues with their anatomical sites noted as colon (in COAD) and rectum and rectosigmoid junction (in READ) were analyzed. i, *WBP11* expression in colon tumors and normal tissues from the TCGA COAD cohort; ii, *WBP11* expression in rectal tumors and normal tissues from the TCGA READ cohort. Expression of *WBP11* is significantly higher in colon and rectum tumors compared to normal tissues. The number of patients in the colon and rectum tumor datasets is larger than those in the normal tissue datasets. This may explain why the gene expression levels in tumors have a higher variance compared to that in the normal tissues.

The three variants with time-varying associations and their high-LD SNPs were also examined, but none of them were found to be eQTLs. Other eQTLs identified in the recessive model analyses are shown in **Supplementary Table 25**.

5.5.4 Examining the associations of high-LD SNP genotypes and *WBP11* expression levels with CMS in the TCGA dataset

A nominal association was detected between rs2241221 and CMS (Fisher's exact test p value = 0.052). Additionally, a significant association was found between *WBP11* expression levels and CMS (Kruskal-Wallis test p value = 9.66×10^{-07}). Pair-wise comparisons further showed that the expression levels of *WBP11* were different between CMS1, CMS2, and CMS4 in the TCGA dataset (**Supplementary Table 26**).

5.5.5 Associations between CNVs/INDELs and survival outcomes

None of the CNVs/INDELs reached the p-value threshold of 5×10^{-06} in the univariate analyses, therefore, were not selected for multivariable analyses. We show the top three CNVs/INDELs identified in the DSS and RMFS analyses in **Supplementary Table 27**.

5.6 Discussion

We investigated the associations of a genome-wide set of common SNPs and 254 CNVs/INDELs with time to death from colorectal cancer (DSS) and time to recurrence/metastasis (RMFS) in a colorectal cancer patient cohort with a long follow-up. As a result, we identified one common SNP, *WBP11*-rs7314075, that was significantly associated with DSS when adjusted for clinical factors (3.27×10^{-08} for dominant model, and 3.24×10^{-08} for additive model). A set of highly linked SNPs with *WBP11*-rs7314075 were also associated with DSS in the TCGA patient cohort. This is one of the first replicated GWAS findings in colorectal cancer. This variant and the SNPs that are in high-LD with them have not been reported in other GWASs [176–179,181,434] and the candidate gene/pathway studies examining the colorectal cancer outcomes (based on the dbCPCO database [451]). Hence, these SNPs are novel candidate prognostic markers in colorectal cancer. In addition, we also identified fifteen genomic loci that were associated with clinical outcomes under the recessive model and they require validation in other independent cohorts. Interestingly, three variants in two genomic loci showed time-varying associations; they predicted the outcome risk after 5 years, but not prior to this time point (i.e. candidate markers of late local/distant recurrent disease). To our knowledge, these variants are the first variants that can predict late recurrent disease in colorectal cancer. On the other hand, in contrast to SNPs, there was no associations of common CNVs/INDELs with the clinical outcomes examined. To our knowledge, it is one of the few GWASs examining colorectal cancer outcomes, the first GWAS that

examines the germline sets of both SNP and CNVs/INDELs in the same patient cohort, and the most comprehensive study examining the time-varying associations of genetic markers with clinical outcomes in colorectal cancer. Overall, with its comprehensive and unique study design, analysis, and results, this study significantly advances the prognostic research in colorectal cancer and expands the knowledge on the relationship of genetic variants with patient outcomes.

5.6.1 Associations with constant HRs (i.e. with proportional hazards)

One common SNP (rs7314075) was associated with DSS under both the dominant and additive genetic models. Further investigations in the TCGA (COAD and READ) patient dataset strengthened our confidence in this association. rs7314075 has a MAF of 14% in the patient cohort and locates in the 8th intron of *WBP11*. *WBP11* encodes a protein that is involved in mRNA splicing [452]. Interestingly, a study on gastric cancer found that inhibiting *WBP11* expression results in the suppression of β-catenin, and thus, suppression of proliferation and migration of tumor cells [453]. β-catenin is a key component of WNT signaling pathway, which is involved in tumorigenesis and disease progress in colorectal cancer [454]. In line with the findings in gastric cancer [453], analysis of the TCGA data showed that the expression levels of *WBP11* in colon and rectum tumors were higher than in adjacent normal tissues. Also, the tumor *WBP11* expression levels were associated with CMS in the TCGA dataset, which is a gene expression-based classification system and has been reported to have associations with outcomes in

colorectal cancer [445]. These findings suggest a possible role of *WBP11* in colorectal cancer that needs to be examined further. According to RegulomeDB [446], there are two SNPs (that are in high-LD with rs7314075) that are annotated as eQTLs in monocytes: rs2241221 and rs11056174. Interestingly, for both eQTLs, the target gene is identified as *FLJ32115/ERP27*. *ERP27* codes for an endoplasmic reticulum protein. An analysis of the TCGA data showed that, similar to *WBP11,* this gene has higher expression levels in colorectal tumors compared to non-tumor samples (**Supplementary Figure 13**). Overall, findings by this study and existing literature suggest a possible biological relationship of *WBP11* with disease outcomes in colorectal cancer, and the *ERP27* gene can be an interest for future studies.

The remaining associations with DSS and RMFS were detected under the recessive genetic model and included variants from three and 10 genomic loci, respectively. While genotypes that are associated with outcomes are relatively rare, these SNPs/loci are worth examining in future studies with larger cohort sizes.

5.6.2 Time-varying associations

Variants in two separate genomic regions were identified to have time-varying associations (i.e. non-proportional hazards) in the RMFS analysis. These genetic markers were able to distinguish between patients with different outcome risk in the long-term (after 5 years post-diagnosis). Minor allele homozygous genotypes of these SNPs predicted shorter RMFS times. According to the gnomAD database [236], the MAF of one of these SNPs (rs817090) is much higher in the African (30%) and Ashkenazi Jewish

(18%) populations than Europeans. Therefore, it is possible that this SNP may predict the outcome risk of a higher number of colorectal cancer patients from these populations. All three variants are located in intergenic regions, and according to RegulomeDB [446] and GTEx [447], there is no strong evidence supporting potential regulatory functions. Similar results were obtained for the SNPs that are in high-LD with them. These findings suggest that further studies are needed to elucidate the biological mechanisms that can explain these SNPs' associations with the recurrent colorectal cancer in the long-term.

Our study significantly contributes to the scientific knowledge on prognostic markers with time-varying associations. This kind of marker is under-studied in colorectal cancer [177,189,196,395]. Since such variables may be missed by traditional analyses, application of appropriate statistical methods, as we have done in this study, is important to detect these variants. Additionally, such markers can provide unique clinical information (e.g. the time-periods of high outcome risk), they can be quite useful in the clinic management of patients. Research into variants with time-varying associations, therefore, should be encouraged. Should the time-varying associations we detected be replicated in independent cohorts, these markers may be used to predict the colorectal cancer patients with a risk of recurrent disease after 5 years. Since clinic surveillance of patients for disease outcomes normally do not continue beyond the first 5 years, such information can be important to identify the patients who have high risk in the long-term. This in turn can facilitate effective surveillance and clinical management of the patients at risk, with an anticipated improvement of their long-term disease outcomes. We hope that our study will inspire more studies specifically looking for this clinically important type of prognostic markers.

5.6.3 Strengths and limitations

This study included common genetic variants, leaving rare variants to be investigated by further studies. We report associations, which are not the same as causation – this should be kept in mind while interpreting our results. We may have missed associations of rare variants and rare genotypes (especially in recessive genetic model analyses), or associations with small effects. Also, while we used a conservative p-value threshold to control type I errors, we cannot rule out the possibility of false-positive findings. Therefore, findings of this study need to be replicated in other colorectal cancer cohorts prior to any clinic utility can be established. In this study, 5 years was chosen as the cut-off time point in survival analysis of the variants that violated the PH assumption. This time point helps define simple and clinically meaningful models. However, there can be variants that have their cut-off time points other than 5 years; such variants can be an interest for future studies. The patient cohort has up to 19 years of follow-up. To our knowledge, this is one of the longest follow-up data in colorectal cancer, which allowed us to examine the time-varying associations, particularly those that appear after the initial 5 years. Also, this study investigated different types of genetic variants (i.e. SNPs, CNVs/INDELs) in the same colorectal cancer cohort. This allowed us to have a comprehensive view of the relationships between genetic variants and survival outcomes in colorectal cancer. In addition, this study assumed no specific genetic model for the tested SNPs, and included analyses under the three main genetic models. Such a comprehensive examination should have limited the possibility of missing SNPs with potential prognostic associations. We also detected the association of a set of SNPs that

are highly linked with *WBP11*-rs7314075 in the TCGA colorectal cancer cohort dataset, increasing our confidence in the association of this SNP with DSS. Last, we made sure that all variables in Cox models satisfied the PH assumption, which increases the reliability of effect inference. More importantly, examining the PH assumption allowed the detection of novel genetic variants with time-varying associations. If validated in independent sets, these markers can help distinguish patients with different outcome risk during select time-periods following diagnosis, and therefore, provide more specific prognostic estimates.

5.7 Conclusions

In conclusion, this study identified a novel common variant (which also showed an association in the TCGA patient dataset) and a number of rare variants, but no CNVs, that are associated with clinical outcomes in colorectal cancer. We also identified genetic variants with time-varying associations, a traditionally under-studied type of prognostic markers. Overall, identified variants/loci - if their prognostic value is validated in independent patient cohorts - can be used to stratify colorectal cancer patients into different risk groups, and help guide the treatment strategies and clinic follow-up in the future.

5.8 Acknowledgments

We would like to express our gratitude for the patients; investigators/staff at NFCCR for collecting and managing the registry data; CHIA for providing computational platform for analyses; and staff at the Provincial Tumor Registry-NL and NLCHI for their help with the clinical data. Yajun Yu was supported by fellowships from TPMI/NLSUPPORT Educational Funding Award; Dean's Fellowship; A. G. Hatcher Memorial Scholarship, and was a trainee in the Cancer Research Trainee Program (CRTP) of the Beatrice Hunter Cancer Research Institute (BHCRI), with funds provided by the Terry Fox Research Institute. Sevtap Savas is a senior scientist of BHCRI. We gratefully acknowledge that the results obtained in the TCGA dataset are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

CHAPTER 6. GENERAL SUMMARY AND DISCUSSION

As one of the most common types of malignancy, colorectal cancer is also one of the leading causes of death among all cancers [1]. Many patients with this disease also experience other outcomes, such as local recurrences and distant metastases. These outcome events greatly burden the patients as well as the health care systems. Predicting risks of these outcome events is very important, as such information can guide the treatment and follow-up strategies for patients and thus extend their survival times. The disease stage and a few other markers are used for such a purpose. However, there is still a need for additional markers for prognostic predictions. Furthermore, the outcome risks of patients may vary over time. Not knowing such patterns complicates disease management for patients. Hence, there is also an urgent need for novel prognostic markers that can predict outcome risks in different time-periods following diagnosis as well. In this work, I examined genetic variants as well as clinico-demographic factors for their potential to be prognostic markers in a colorectal cancer cohort from Newfoundland and Labrador. A number of factors were identified to have associations, as well as time-varying associations, with survival outcomes of patients.

Two INDELs in the promoter region of the *BRM* gene were first examined in colorectal cancer using the NFCCR cohort followed up to around 11 years (Chapter 2). Studies have shown that these two INDELs can affect the expression of *BRM* [95,279,323]. *BRM* is a gene involved in chromatin remodeling and thus can influence the expression of many genes, including genes that function in signal transduction, cell cycle, DNA repair, and others that link to cell proliferation, differentiation, invasion, and migration [455–459]. The two *BRM* INDELs thus are interesting, as they may affect important biological processes (including processes that influence cancer

prognosis) through their impacts on the *BRM* expression. What makes these two INDELs more interesting is that they have already been reported to be involved in prognoses of multiple cancers (e.g., lung [325], liver [326], esophageal [327], pancreatic [324] cancers, and malignant pleural mesothelioma [99]). However, their prognostic value in colorectal cancer prior to my research was not known. The *BRM* study (Chapter 2) of my research identified the *BRM*-741 to have an association (with no evidence of time-varying associations) with progression-free survival in colorectal cancer, though its association pattern was different than in other studies [99,324–327]. In those studies, heterozygotes (Ins/Del) and homozygotes with double variant alleles (Ins/Ins) of *BRM*-741 usually had worse outcomes compared to homozygotes with double wild-type alleles (Del/Del). However, in my study, the heterozygotes (Ins/Del) of this variant had better outcomes, and homozygotes with double variant alleles (Ins/Ins) and with double wild-type alleles (Del/Del) had no significant difference in their survival times. The reason for this is unknown yet and warrants further investigation. Nevertheless, my study showed that there is an association of this *BRM* variant with colorectal cancer prognosis, and this implicates a plausible role of this variant (and its related gene, the *BRM*) in tumor progression in colorectal cancer, in addition to other cancers (lung, liver, esophageal, pancreatic cancers and malignant pleural mesothelioma). As part of the study, associations of *BRM*-741 and *BRM*-1321 with the disease risk/susceptibility of colorectal cancer were also examined, and it was found that at least one variant allele (Ins) in either of these two INDELs was associated with an increased risk of colon cancer but not rectal cancer. Overall, this study contributes to the rapidly growing scientific knowledge on these two interesting INDELs in the *BRM* gene, and the findings may also inspire other groups.

The prognostic value of the majority of germline CNVs and INDELs in the human genome is unknown in colorectal cancer. The genic CNVs/INDELs study (Chapter 3) examined 106 common genic CNVs/INDELs that overlapped with the genes in the human genome for their associations with the relapse-free survival in colorectal cancer. A number of variants (five CNVs in *TGFBR3*, *STEAP2*, *RP11-143P4.2*, *PDLIM3*, *GUSBP1*, and an INDEL in *FILIP1L*) were identified to have associations. Interestingly, two of the identified CNVs (*PDLIM3* and *GUSBP1* CNVs) were novel variants to have time-varying associations; and therefore, they are candidate early-relapse markers (as their associations were only detected within the first ~3 years post-diagnosis). In this study, results of the survival analyses with and without checking the PH assumption were also compared. As a result, it was clear that variants with time-varying associations (*PDLIM3* and *GUSBP1* CNVs) cannot be detected when the PH assumption was not checked (i.e., assuming this assumption was satisfied) and when only the conventional Cox PH model was used. However, they can be identified when the PH assumption was checked and using appropriate survival models, as I have done in this study. This underscores the importance of checking the PH assumption and using appropriate models in survival analyses, similar to other studies [177,185,190–196]. Moreover, considering that only two out of 106 CNVs/INDELs were variants with non-constant associations, this study also suggested that there might not be many common CNVs/INDELs in the human genome that have time-varying associations with survival outcomes in colorectal cancer.

While the *BRM* and 106 genic CNVs/INDELs studies were based on the NFCCR cohort followed up to around 11 years, my third study (cli210nico-demographic factors study; Chapter 4) was based on the data of the NFCCR cohort followed up to 19 years. This study described the

prognostic characteristics of this long follow-up data, which is one of the longest ever documented in the literature in colorectal cancer. While the majority of the disease-related outcome events (e.g., recurrence and metastasis) occurred within the first years (up to 6 years post-diagnosis), there were still around 15% of such events detected after that. This may be informative for clinicians to decide whether colorectal cancer patients should be followed longer than 5 years after their diagnosis (the 5-year follow-up period is generally suggested by follow-up guidelines (**Table 1.2**)). More importantly, this study examined the associations, especially the time-varying associations, of cli211nico-demographic factors with a detailed set of outcomes (i.e., OS, DSS, RFS, MFS, RMFS, and EFS). Different sets of factors were found to be associated with different outcomes. A number of factors (e.g., MSI status) had constant associations over time, and they can predict the corresponding outcomes till 19 years after diagnosis. Intriguingly, some other factors were found to have time-varying associations (e.g., *BRAF* Val600Glu mutation and tumor location). They are thus candidate markers that can predict outcome risks at different times along the survival times. With regard to the mechanisms of these time-varying associations, many remain unknown and need to be further investigated. Overall, this study described the long-term prognostic characteristics of a colorectal cancer cohort that was followed up to 19 years, and for the first time, described a comprehensive picture of the associations, including time-varying associations, of cli211nico-demographic factors with different outcomes in colorectal cancer.

Using the updated follow-up data of the NFCCR cohort, a genome-wide set of common SNPs (n = ~4.7 million) were further examined for their associations with select clinical outcomes in colorectal cancer (Chapter 5). In addition, 254 common CNVs/INDELs, which are

located in both genic and intergenic regions across the genome, were also examined along with the genome-wide SNP dataset. This study focused on two clinically important outcomes, disease-specific survival (DSS) and recurrence/metastasis-free survival (RMFS), which are different than that of the other genetic association studies I conducted (Chapters 2 and 3). A SNP in *WBP11* was identified to have an association with DSS, and this association was further supported by analyzing the TCGA data, an independent cohort data from the USA. Gene expression analysis using bioinformatics resources showed that the expression level of *WBP11* was different between tumors and adjacent normal tissues, further suggesting a role of *WBP11* in colorectal cancer prognosis. High-LD SNPs of the *WBP11*-SNP were also found to be eQTLs in monocytes (associated gene: *ERP27*), providing other information on genes that are interesting for future investigation. In addition, three SNPs in two different genomic regions (on chromosomes 2 and 12) were identified to have associations with RMFS only after 5 years post-diagnosis (time-varying associations). If this finding is replicated, they can be the first examples of genetic late-recurrence/metastasis markers in colorectal cancer. Overall, this study showed that SNPs, but not CNVs/INDELs, have associations with DSS and RMFS in colorectal cancer, and it also provided a number of SNPs (and genes) that may have a role in colorectal cancer prognosis.

This thesis research is one of the few studies that comprehensively analyzed major genetic variations (SNPs, CNVs/INDELs) for their potential impacts on the prognosis of colorectal cancer, providing insights into the potential prognostic roles of major genetic variations in this disease. Although SNPs are the most common type of genetic variations in the human genome and the most investigated variants in large-scale studies, in colorectal cancer, there were only six GWASs (**Table 1.4**) [176–181] performed for their associations with prognosis

prior to my study (Chapter 5). These studies identified only 16 SNPs that were associated with OS, PFS, MFS, or DSS. Except for OS, there was only one GWAS performed for each of these outcomes, indicating more studies on these (and other uninvestigated) outcomes are needed. In the GWAS of this thesis research (Chapter 5), as mentioned earlier, SNPs were investigated for their associations with RMFS and DSS, making the study the first and second GWAS to examine these two outcomes, respectively. In addition, the GWAS of this research is also the first GWAS that examined SNPs and CNVs/INDELs from the same patient cohort in the same study. The identified SNPs were novel variants that have not been reported in other studies conducted on both disease prognosis and susceptibility in colorectal cancer [176–181,451,460]. Compared to SNPs, CNVs and INDELs received much less attention in large-scale studies on disease prognosis, partially because it is harder to map/annotate structural variations in the genome than SNPs [308]. Indeed, structural variants affect more nucleotides than SNPs in a given genome [235], whereas their prognostic roles in colorectal cancers are far from understood. Amongst the studies that investigated CNVs/INDELs in this disease, many focused on somatic CNVs and INDELs [461–463]. The 106 genic CNVs/INDELs study (Chapter 3) was the first study that systematically investigated germline CNVs/INDELs for their associations with survival outcomes (relapse-free survival) in colorectal cancer, though only the genic ones were examined. Both the genic and intergenic CNVs/INDELs were investigated in the GWAS study (Chapter 5). Together with the investigation on two INDELs in a candidate gene *BRM* (Chapter 2), this thesis research is pioneer research presenting a comprehensive view of relationships between CNVs/INDELs and different outcomes in colorectal cancer.

It has been reported that common CNVs/INDELs may be well-represented by nearby common SNPs (i.e., in high LD with each other) [285,359]; hence, associations of common CNVs/INDELs with diseases, if there is any, may have already been indirectly screened by SNP-based studies [285,359]. Though this cannot be verified in the GWAS (Chapter 5) of this research as no CNVs/INDELs were identified in that study, other studies (Chapters 2 and 3) of this thesis research found some common CNVs/INDELs to be associated with disease prognosis (*BRM*-741 and *BRM*-1321 INDELs on chromosome 9; *PDLIM3* and *GUSBP1* CNVs on chromosomes 4 and 5, respectively). However, in previous GWASs [176–179,181,434] on colorectal cancer prognosis (though the outcome measures might be different for some GWASs), no identified SNPs are located close to (or are not even on the same chromosomes with) these CNVs/INDELs. The CNVs/INDELs identified in this thesis research are thus CNVs/INDELs with independent associations (i.e., associations that are not attributed to SNPs in high-LD with them). This indicates that associations of common CNVs/INDELs may not always be well-represented by SNPs and captured by SNP-based studies. This is not a surprising finding, as around 23% of the common CNVs are not in high-LD with nearby SNPs [285,359]. Further, considering the fact that many CNVs/INDELs are likely to be pathogenic [263,464], including these variations in association analysis increases the chance of pinpointing the causal variants. Also, in SNP-based studies, if resolutions of genotyping platforms are not high, identifying variants with associations can be (severely) affected [465], especially for those variants that are non-SNP variants (e.g., CNVs/INDELs). Thus, examining common CNVs/INDELs is not redundant but necessary in dissecting variants that have impacts on diseases. Of course, rare CNVs/INDELs, similar to rare SNPs, deserve to be investigated in future studies, as they are even less investigated and believed to account for parts of the "missing heritability" in cancer risks and prognoses [25,466].

This research is also the first study that systematically investigated the relationships between cli215nico-demographic factors with a detailed set of outcomes in colorectal cancer using a long follow-up data. Clinico-demographic factors are commonly investigated factors for their potential as prognostic markers. However, these factors have not been systematically examined for their associations, particularly time-varying associations, with a detailed set of outcomes using a long follow-up data in colorectal cancer, as I have done in Chapter 4. As shown in **Table 1.5**, time-varying associations of cli215nico-demographic factors were previously reported only in analyses for OS and DSS. This thesis research, for the first time, examined time-varying associations of these factors with also other outcomes (RFS, MFS, RMFS, and EFS), and found a number of novel time-varying associations (e.g., tumor location on RMFS). As part of the study, time-varying associations of cli215nico-demographic factors with OS and DSS were also checked, and non-constant associations of age at diagnosis, tumor location, and disease stage with these outcomes were detected in our research as well, strengthening our confidence in their time-varying associations with OS and DSS in colorectal cancer [190–195]. Clinico-demographic factors with no time-varying associations have long-term and constant associations with outcomes in this disease. Indeed, some of these factors have already been known to have prognostic value; however, if not being checked for their PH assumptions and analyzed in a long follow-up data, we could never be sure that their associations were constant over a long period of time. These findings, therefore, significantly add up to the knowledge of colorectal cancer prognosis and its relation to cli215nico-demographic factors.

Regarding the prognostic markers with time-varying associations specifically, this thesis research also adds significantly to the field. Such markers are a specific type of prognostic

marker, as they can predict the patients' early or late outcomes and may further help optimize the treatment and management strategies for patients. Though important, prognostic markers with time-varying associations are largely under-studied in cancer research. Among the limited number of studies investigating such markers, most focused on breast cancer [185,216,393,467–479]. Only a few studies were performed in colorectal cancer (**Table 1.5**). What is more, these studies mainly investigated clinico-demographic factors, not genetic variations. My research work thus filled this gap by examining large-scale genetic variants for their potential to be prognostic markers with time-varying associations, using the data from a colorectal cancer cohort (i.e., NFCCR cohort) followed up for a long time. The long follow-up data of the cohort not only increased the study power but also elevated the chance of identifying factors with such associations, especially those associated with late outcomes in colorectal cancer. This thesis research successfully identified some variants that have the potential to be prognostic markers with time-varying associations (Chapters 3 and 5). Other than that, a number of clinico-demographic factors that were deemed to have constant associations were actually identified as candidate prognostic markers with time-varying associations. For example, tumor location is generally believed to have a constant association with OS [480,481]; however, in this research (Chapter 4), I showed that rectal tumor was associated with an increased risk of overall death only after 2 years post-diagnosis, but not before that. Age at diagnosis is another factor that is generally regarded to have constant associations with outcomes, but my research (Chapters 2-4) showed that its associations with outcomes changed over time, possibly due to the aging of the study population over time. Similarly, disease stage and adjuvant chemotherapy also showed rarely reported time-varying associations with outcomes in my research (Chapter 4). Associations of disease stages III and IV with death-related outcomes were diminished after 1-

1.5 years post-diagnosis, which may be due to advanced disease at diagnosis and/or the post-surgical complications that usually lead to early death [407–409]. Adjuvant chemotherapy was found to have strong associations with death-related outcomes in the first-year post-diagnosis but weaker/disappeared associations after that in colorectal cancer, possibly reflecting the diminishing effects of drugs over time. Other than the interesting findings of having time-varying associations for some clinico-demographic factors that are generally believed to have constant associations, this research also found a clinical factor (i.e., adjuvant radiotherapy) that is deemed to be protective in prognosis may have a detrimental effect on metastasis at late years post-diagnosis. Patients who received the adjuvant radiotherapy treatment had an elevated risk of metastasis after 5.5 years post-diagnosis, but not before that (Chapter 4). Though it is not clear yet about the exact mechanism of this late association, radiation has been reported to be able to increase the risk of tumor metastasis [422–424], possibly through mechanisms related to appearance or development of radiation-resistant tumor cells, change of the tumor microenvironment and immune system response, and others [422–424]. My findings and other published literature thus provide insights into the prognostic effect of radiotherapy treatment over time.

This thesis research further highlights the importance of checking the PH assumption in survival studies using the Cox model. As described earlier (Section 1.7.1.1), checking the assumption of proportional hazards is not a common practice in medical research. Jachno and colleagues checked 66 clinical trials with time-to-event as the primary outcome and found that ~89% of studies just assumed constant associations for all examined factors [183]. Carroll and others found that 72% of 148 survival studies in oncology did not check/report the PH assumption [182]. As described by others [183,185], because the assumption was not checked,

inappropriate models might be used, and hence the findings can be misleading. Furthermore, factors with time-varying associations can be missed by those studies. This was affirmed by the 106 genic CNVs/INDELs study in which results of models based on checking and not checking the PH assumption were compared (Chapter 3). The two variants (*PDLIM3* and *GUSBP1* CNVs) with time-vary associations were only detected in models under the setting where the PH assumption was checked. Actually, one of the most valuable parts of this thesis research is that the PH assumption was checked for all factors examined, and proper models were used to examine the associations between factors and outcomes. The results are thus considered more reliable compared to many other survival studies in medical research. Other than that, by taking advantage of checking the PH assumption, this thesis research also identified factors with time-varying associations. These underscore the importance of checking the PH assumption of the Cox model and using proper models for prognostic investigations to create reliable as well as novel knowledge.

The piecewise/change-point Cox PH model was used in this research when the PH assumption was violated. This model is simple, straightforward, and clinically meaningful to clinicians, as time-varying associations were simply described as different associations before and after specific cut-off time points. Other models can also be used when the PH assumption is violated, but usually require the prior knowledge of the survival/hazard distributions (e.g., parametric models and related models), allow the associations to change at every single time point (which made the model a more complex model; e.g., Cox PH model with time-varying coefficients [coefficients as continuous functions of time] and the Aalen's additive model), or may even not be able to estimate the time-varying associations at all (e.g., stratified Cox PH

model). However, this does not mean that these models are not good options. On the contrary, they are suitable alternatives to the Cox PH model under certain conditions. For example, when the factor that violates the PH assumption is not the factor of interest, the stratified Cox PH model can be a good option [210,211]. This is because, in such a case, no estimation of the association for this factor is acceptable. Another example is that when the survival/hazard distribution is known based on considerable prior knowledge and it makes the related model has a good fit to data, the parametric model can be a good alternative for factors that violate the PH assumption, as they usually give more precise effect estimations compared to semi-parametric models [482,483].

My research consisting of four different published studies depicted a detailed picture of the common genetic variants and clinico-demographic factors with and without time-varying associations in colorectal cancer. Through these studies, it can be seen that: (1) the majority of common genetic variants (both SNPs and CNVs/INDELs) do not seem to have associations, including time-varying associations, with survival outcomes in colorectal cancer; (2) there are genetic variants and baseline clinico-demographic factors that have long-term and constant associations in colorectal cancer (e.g., two *BRM* INDELs, four CNVs/INDELs in the 106 genic CNV/INDEL study, the MSI status, and *WBP11*-rs7314075), and they may predict the outcome risks even long after diagnosis; (3) a number of genetic variants and clinico-demographic factors also have time-varying associations with disease outcomes (e.g., two CNVs and three SNPs), and they are among the first candidate genetic early- or late-outcome markers; (4) checking the PH assumption is very important and should be encouraged in survival studies, as this is pivotal to obtain more reliable results and identify factors with time-varying associations; (5) patients

survived the first ~6 years tend to have a low chance to experience disease-related outcomes. However, there are still around 15% of first recurrences and metastases that occur after that, raising the question of whether the most recommended 5 years follow-up frame should be extended or not.

This research also has a number of "firsts": (1) it is the first research that two *BRM* INDELs and large sets of germline CNVs/INDELs have been examined for their associations, including time-varying associations, with survival outcomes in colorectal cancer; (2) it is the first research where the long-term survival characteristics and the associations of baseline factors with a comprehensive set of clinical outcomes in colorectal cancer were examined/identified, by taking advantage of a long follow-up data; (3) it is the first research that both a large number of SNPs and CNVs/INDELs from the same patient cohort were examined in relation to survival outcomes of colorectal cancer patients, making it the most comprehensive genetic survival study in colorectal cancer; (4) it is also the first GWAS that investigated RMFS (the second GWAS for DSS) in colorectal cancer. Also, my research study is one of the first research studies that investigated the time-varying associations of genetic variants in colorectal cancer, as well as one of the first GWASs on prognoses of colorectal cancer patients.

6.1 Future directions

Following this research, the identified novel associations, including the time-varying associations, warrant further validations in other large datasets before any factors can be used in clinics as prognostic markers in colorectal cancer. Other than the time-varying associations of

genetic variants identified in the Chapter 5 study, additional time-varying associations of these variants may also be identified using cut-off time points other than 5 years post-diagnosis; this can be examined in the NFCCR and/or other colorectal cancer cohorts in the future. In addition, the mechanisms of time-varying associations are unclear and warrant further investigation, although such associations of some clinical factors have possible explanations (e.g., adjuvant chemotherapy - changed associations with death-related outcomes after year 1 post-diagnosis may reflect the diminishing effect of the drug over time; adjuvant radiotherapy – changed association with metastasis after 5.5 years post-diagnosis may be due to the appearance or development of radiation-resistant tumor cells, change of the tumor microenvironment and immune system response over time, and others; age at diagnosis – associations change over time may be attributed to the aging of the study population over time; and disease stages III and IV – diminishing associations after 1-1.5 years post-diagnosis may be caused by advanced disease at diagnosis and/or the post-surgical complications that lead to early death) (Chapter 4). Uncovering the mechanisms of time-varying associations, in return, increases our confidence in utilizing related factors as prognostic markers. Also, checking whether the identified factors have causal effects on prognosis can be an interest of future studies. To do that, investigations on related variants/genes/factors and their function/role in cell lines and/or animal models, as well as other necessary experimental and bioinformatical analyses, need to be performed. Indeed, some factors may have no direct causal links to prognosis, and they may merely reflect the effects of causal factors (e.g., SNPs that are in high-LD with the causal SNPs). In such cases, interested researchers can perform studies to fine map the causal factors. With regard to the factors investigated in this research, genetic variants were only examined if they were common in the cohort, leaving other variants to be investigated in future studies. Similarly, not all clinical

factors were examined in the research, and those that were not included can be analyzed in the future. Besides, interactions among identified genetic variants and between genetic variants and environmental factors can be other interests of future studies.

In conclusion, this thesis research identified a number of genetic variants (SNPs and CNVs/INDELs) that have associations with survival outcomes in colorectal cancer. In addition, different clinico-demographic factors were proven to have prognostic value in various outcomes in this disease. Of the examined genetic variants and clinico-demographic factors, some were found to have time-varying associations with outcomes of patients. Thus, they are candidate early- or late-outcome markers. If these findings are validated, the factors identified in this research can help stratify patients into different outcome-risk groups, and guide clinicians to design appropriate treatment strategies for patients. These, in the end, may help reduce the disease burdens of colorectal cancer and extend the survival times of patients.

BIBLIOGRAPHY

1. Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: incidence, mortality, survival, and risk factors. *Przeglad Gastroenterol.* **14**, 89–103 (2019) doi:10.5114/pg.2018.81072.
2. Torre, L. A. *et al.* Global cancer statistics, 2012. *CA. Cancer J. Clin.* **65**, 87–108 (2015) doi:10.3322/caac.21262.
3. Ferlay, J. *et al. Global Cancer Observatory: Cancer Today*. https://gco.iarc.fr/today (2020).
4. Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* (2021) doi:10.3322/caac.21660.
5. Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA. Cancer J. Clin.* **55**, 74–108 (2005) doi:10.3322/canjclin.55.2.74.
6. Torre, L. A., Siegel, R. L., Ward, E. M. & Jemal, A. Global Cancer Incidence and Mortality Rates and Trends--An Update. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **25**, 16–27 (2016) doi:10.1158/1055-9965.EPI-15-0578.
7. Aran, V., Victorino, A. P., Thuler, L. C. & Ferreira, C. G. Colorectal Cancer: Epidemiology, Disease Mechanisms and Interventions to Reduce Onset and Mortality. *Clin. Colorectal Cancer* **15**, 195–203 (2016) doi:10.1016/j.clcc.2016.02.008.
8. Arnold, M. *et al.* Global patterns and trends in colorectal cancer incidence and mortality. *Gut* **66**, 683–691 (2017) doi:10.1136/gutjnl-2015-310912.

9. Canadian Cancer Statistics Advisory Committee. *Canadian Cancer Statistics 2019*. Available at: https://cdn.cancer.ca/-/media/files/research/cancer-statistics/2019-statistics/canadian-cancer-statistics-2019-en.pdf (accessed on Nov 23, 2021).

10. Nguyen, H. T. & Duong, H.-Q. The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy. *Oncol. Lett.* **16**, 9–18 (2018) doi:10.3892/ol.2018.8679.

11. Souglakos, J. Genetic alterations in sporadic and hereditary colorectal cancer: implementations for screening and follow-up. *Dig. Dis. Basel Switz.* **25**, 9–19 (2007) doi:10.1159/000099166.

12. Giglia, M. D. & Chu, D. I. Familial Colorectal Cancer: Understanding the Alphabet Soup. *Clin. Colon Rectal Surg.* **29**, 185–195 (2016) doi:10.1055/s-0036-1584290.

13. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990) doi:10.1016/0092-8674(90)90186-i.

14. Grady, W. M. Genetic testing for high-risk colon cancer patients. *Gastroenterology* **124**, 1574–1594 (2003) doi:10.1016/s0016-5085(03)00376-7.

15. Jasperson, K. W., Tuohy, T. M., Neklason, D. W. & Burt, R. W. Hereditary and familial colon cancer. *Gastroenterology* **138**, 2044–2058 (2010) doi:10.1053/j.gastro.2010.01.054.

16. Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N. Engl. J. Med.* **343**, 78–85 (2000) doi: 10.1056/NEJM200007133430201.

17. Wells, K. & Wise, P. E. Hereditary Colorectal Cancer Syndromes. *Surg. Clin. North Am.* **97**, 605–625 (2017) doi:10.1016/j.suc.2017.01.009.

18. Chen, E., Xu, X. & Liu, T. Hereditary Nonpolyposis Colorectal Cancer and Cancer Syndromes: Recent Basic and Clinical Discoveries. *J. Oncol.* **2018**, 3979135 (2018) doi:10.1155/2018/3979135.

19. Dinarvand, P. *et al.* Familial Adenomatous Polyposis Syndrome: An Update and Review of Extraintestinal Manifestations. *Arch. Pathol. Lab. Med.* **143**, 1382–1398 (2019) doi:10.5858/arpa.2018-0570-RA.

20. Testa, U., Pelosi, E. & Castelli, G. Colorectal Cancer: Genetic Abnormalities, Tumor Progression, Tumor Heterogeneity, Clonal Evolution and Tumor-Initiating Cells. *Med. Sci.* **6**, 31 (2018) doi:10.3390/medsci6020031.

21. Schatoff, E. M., Leach, B. I. & Dow, L. E. Wnt Signaling and Colorectal Cancer. *Curr. Colorectal Cancer Rep.* **13**, 101–110 (2017) doi:10.1007/s11888-017-0354-9.

22. Kantor, M., Sobrado, J., Patel, S., Eiseler, S. & Ochner, C. Hereditary Colorectal Tumors: A Literature Review on MUTYH-Associated Polyposis. *Gastroenterol. Res. Pract.* **2017**, 8693182 (2017) doi:10.1155/2017/8693182.

23. Sampson, J. R. *et al.* Autosomal recessive colorectal adenomatous polyposis due to inherited mutations of MYH. *Lancet Lond. Engl.* **362**, 39–41 (2003) doi:10.1016/S0140-6736(03)13805-6.

24. Vadde, R. *et al.* Role of hypoxia-inducible factors (HIF) in the maintenance of stemness and malignancy of colorectal cancer. *Crit. Rev. Oncol. Hematol.* **113**, 22–27 (2017) doi:10.1016/j.critrevonc.2017.02.025.

25. Peters, U., Bien, S. & Zubair, N. Genetic architecture of colorectal cancer. *Gut* **64**, 1623–1636 (2015) doi:10.1136/gutjnl-2013-306705.

26. Lindor, N. M. Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome. *Surg. Oncol. Clin. N. Am.* **18**, 637–645 (2009) doi:10.1016/j.soc.2009.07.003.

27. Nejadtaghi, M., Jafari, H., Farrokhi, E. & Samani, K. G. Familial Colorectal Cancer Type X (FCCTX) and the correlation with various genes-A systematic review. *Curr. Probl. Cancer* **41**, 388–397 (2017) doi:10.1016/j.currproblcancer.2017.10.002.

28. Armelao, F. & de Pretis, G. Familial colorectal cancer: a review. *World J. Gastroenterol.* **20**, 9292–9298 (2014) doi:10.3748/wjg.v20.i28.9292.

29. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**, 695–701 (2008) doi:10.1038/ng.f.136.

30. de la Chapelle, A. Genetic predisposition to colorectal cancer. *Nat. Rev. Cancer* **4**, 769–780 (2004) doi:10.1038/nrc1453.

31. Bogaert, J. & Prenen, H. Molecular genetics of colorectal cancer. *Ann. Gastroenterol.* **27**, 9–14 (2014). PMID: 24714764.

32. Yamagishi, H., Kuroda, H., Imai, Y. & Hiraishi, H. Molecular pathogenesis of sporadic colorectal cancers. *Chin. J. Cancer* **35**, 4 (2016) doi:10.1186/s40880-015-0066-y.

33. Kim, T. J., Kim, E. R., Hong, S. N., Chang, D. K. & Kim, Y.-H. Long-Term Outcome and Prognostic Factors of Sporadic Colorectal Cancer in Young Patients: A Large Institutional-Based Retrospective Study. *Medicine (Baltimore)* **95**, e3641 (2016) doi:10.1097/MD.0000000000003641.

34. Mundade, R., Imperiale, T. F., Prabhu, L., Loehrer, P. J. & Lu, T. Genetic pathways, prevention, and treatment of sporadic colorectal cancer. *Oncoscience* **1**, 400–406 (2014) doi:10.18632/oncoscience.59.

35. Carethers, J. M. & Jung, B. H. Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer. *Gastroenterology* **149**, 1177-1190.e3 (2015) doi:10.1053/j.gastro.2015.06.047.

36. Carr, P. R. *et al.* Lifestyle factors and risk of sporadic colorectal cancer by microsatellite instability status: a systematic review and meta-analyses. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **29**, 825–834 (2018) doi:10.1093/annonc/mdy059.

37. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012) doi:10.1038/nature11252.

38. Al-Sohaily, S., Biankin, A., Leong, R., Kohonen-Corish, M. & Warusavitarne, J. Molecular pathways in colorectal cancer. *J. Gastroenterol. Hepatol.* **27**, 1423–1431 (2012) doi:10.1111/j.1440-1746.2012.07200.x.

39. Tariq, K. & Ghias, K. Colorectal cancer carcinogenesis: a review of mechanisms. *Cancer Biol. Med.* **13**, 120–135 (2016) doi:10.28092/j.issn.2095-3941.2015.0103.

40. Pino, M. S. & Chung, D. C. The chromosomal instability pathway in colon cancer. *Gastroenterology* **138**, 2059–2072 (2010) doi:10.1053/j.gastro.2009.12.065.

41. McClelland, S. E. Role of chromosomal instability in cancer progression. *Endocr. Relat. Cancer* **24**, T23–T31 (2017) doi:10.1530/ERC-17-0187.

42. Venkatesan, S., Natarajan, A. T. & Hande, M. P. Chromosomal instability--mechanisms and consequences. *Mutat. Res. Genet. Toxicol. Environ. Mutagen.* **793**, 176–184 (2015) doi:10.1016/j.mrgentox.2015.08.008.

43. Oliveira, E. J., Pádua, J. G., Zucchi, M. I., Vencovsky, R. & Vieira, M. L. C. Origin, evolution and genome distribution of microsatellites. *Genet. Mol. Biol.* **29**, 294–307 (2006) doi:10.1590/S1415-47572006000200018.

44. Boland, C. R. *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.* **58**, 5248–5257 (1998) PMID: 9823339.

45. Fleming, M., Ravula, S., Tatishchev, S. F. & Wang, H. L. Colorectal carcinoma: Pathologic aspects. *J. Gastrointest. Oncol.* **3**, 153–173 (2012) doi:10.3978/j.issn.2078-6891.2012.030.

46. Kang, S. *et al.* The significance of microsatellite instability in colorectal cancer after controlling for clinicopathological factors. *Medicine (Baltimore)* **97**, e0019 (2018) doi:10.1097/MD.0000000000010019.

47. Copija, A., Waniczek, D., Witkoś, A., Walkiewicz, K. & Nowakowska-Zajdel, E. Clinical Significance and Prognostic Relevance of Microsatellite Instability in Sporadic Colorectal Cancer Patients. *Int. J. Mol. Sci.* **18**, E107 (2017) doi:10.3390/ijms18010107.

48. Battaglin, F., Naseem, M., Lenz, H.-J. & Salem, M. E. Microsatellite instability in colorectal cancer: overview of its clinical significance and novel perspectives. *Clin. Adv. Hematol. Oncol. HO* **16**, 735–745 (2018) PMID: 30543589.

49. Kloor, M., Staffa, L., Ahadova, A. & von Knebel Doeberitz, M. Clinical significance of microsatellite instability in colorectal cancer. *Langenbecks Arch. Surg.* **399**, 23–31 (2014) doi:10.1007/s00423-013-1112-3.

50. Nazemalhosseini Mojarad, E., Kuppen, P. J., Aghdaei, H. A. & Zali, M. R. The CpG island methylator phenotype (CIMP) in colorectal cancer. *Gastroenterol. Hepatol. Bed Bench* **6**, 120–128 (2013) PMID: 24834258.

51. Rhee, Y.-Y., Kim, K.-J. & Kang, G. H. CpG Island Methylator Phenotype-High Colorectal Cancers and Their Prognostic Implications and Relationships with the Serrated Neoplasia Pathway. *Gut Liver* **11**, 38–46 (2017) doi:10.5009/gnl15535.

52. Barault, L. *et al.* Hypermethylator phenotype in sporadic colon cancer: study on a population-based series of 582 cases. *Cancer Res.* **68**, 8541–8546 (2008) doi:10.1158/0008-5472.CAN-08-1171.

53. Hawkins, N. *et al.* CpG island methylation in sporadic colorectal cancers and its relationship to microsatellite instability. *Gastroenterology* **122**, 1376–1387 (2002) doi:10.1053/gast.2002.32997.

54. Issa, J.-P. CpG island methylator phenotype in cancer. *Nat. Rev. Cancer* **4**, 988–993 (2004) doi:10.1038/nrc1507.

55. Shen, L. *et al.* Association between DNA methylation and shortened survival in patients with advanced colorectal cancer treated with 5-fluorouracil based chemotherapy. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **13**, 6093–6098 (2007) doi:10.1158/1078-0432.CCR-07-1011.

56. Ward, R. L. *et al.* Adverse prognostic effect of methylation in colorectal cancer is reversed by microsatellite instability. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **21**, 3729–3736 (2003) doi:10.1200/JCO.2003.03.123.

57. Jass, J. R. Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology* **50**, 113–130 (2007) doi:10.1111/j.1365-2559.2006.02549.x.

58. Issa, J.-P. J., Shen, L. & Toyota, M. CIMP, at Last. *Gastroenterology* **129**, 1121–1124 (2005) doi:10.1053/j.gastro.2005.07.040.

59. Bosman, F., Carneiro, F., Hruban, R. & Theise, N (eds). *WHO classification of tumours of the digestive system*. (International Agency for Research on Cancer, 2010).

60. Yamane, L., Scapulatempo-Neto, C., Reis, R. M. & Guimarães, D. P. Serrated pathway in colorectal carcinogenesis. *World J. Gastroenterol.* **20**, 2634–2640 (2014) doi:10.3748/wjg.v20.i10.2634.

61. De Palma, F. *et al.* The Molecular Hallmarks of the Serrated Pathway in Colorectal Cancer. *Cancers* **11**, 1017 (2019) doi:10.3390/cancers11071017.

62. Bevan, R. & Rutter, M. D. Colorectal Cancer Screening-Who, How, and When? *Clin. Endosc.* **51**, 37–49 (2018) doi:10.5946/ce.2017.141.

63. Hadjipetrou, A., Anyfantakis, D., Galanakis, C. G., Kastanakis, M. & Kastanakis, S. Colorectal cancer, screening and primary care: A mini literature review. *World J. Gastroenterol.* **23**, 6049–6058 (2017) doi:10.3748/wjg.v23.i33.6049.

64. Lee, S.-H., Park, Y.-K., Lee, D.-J. & Kim, K.-M. Colonoscopy procedural skills and training for new beginners. *World J. Gastroenterol.* **20**, 16984–16995 (2014) doi:10.3748/wjg.v20.i45.16984.

65. Ahmed, S., Johnson, K., Ahmed, O. & Iqbal, N. Advances in the management of colorectal cancer: from biology to treatment. *Int. J. Colorectal Dis.* **29**, 1031–1042 (2014) doi:10.1007/s00384-014-1928-5.

66. Stintzing, S. Management of colorectal cancer. *F1000prime Rep.* **6**, 108 (2014) doi:10.12703/P6-108.

67. Chakedis, J. & Schmidt, C. R. Surgical Treatment of Metastatic Colorectal Cancer. *Surg. Oncol. Clin. N. Am.* **27**, 377–399 (2018) doi:10.1016/j.soc.2017.11.010.

68. Vodenkova, S. *et al.* 5-fluorouracil and other fluoropyrimidines in colorectal cancer: Past, present and future. *Pharmacol. Ther.* **206**, 107447 (2020) doi:10.1016/j.pharmthera.2019.107447.

69. Vértessy, B. G. & Tóth, J. Keeping uracil out of DNA: physiological role, structure and catalytic mechanism of dUTPases. *Acc. Chem. Res.* **42**, 97–106 (2009) doi:10.1021/ar800114w.

70. Rose, M. G., Farrell, M. P. & Schmitz, J. C. Thymidylate Synthase: A Critical Target for Cancer Chemotherapy. *Clin. Colorectal Cancer* **1**, 220–229 (2002) doi:10.3816/CCC.2002.n.003.

71. Vanhoefer, U. *et al.* Irinotecan in the treatment of colorectal cancer: clinical overview. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **19**, 1501–1518 (2001) doi:10.1200/JCO.2001.19.5.1501.

72. Culy, C. R., Clemett, D. & Wiseman, L. R. Oxaliplatin. A review of its pharmacological properties and clinical efficacy in metastatic colorectal cancer and its potential in other malignancies. *Drugs* **60**, 895–924 (2000) doi:10.2165/00003495-200060040-00005.

73. Xie, Y.-H., Chen, Y.-X. & Fang, J.-Y. Comprehensive review of targeted therapy for colorectal cancer. *Signal Transduct. Target. Ther.* **5**, 22 (2020) doi:10.1038/s41392-020-0116-z.

74. Feeney, G. *et al.* Neoadjuvant radiotherapy for rectal cancer management. *World J. Gastroenterol.* **25**, 4850–4869 (2019) doi:10.3748/wjg.v25.i33.4850.

75. Häfner, M. F. & Debus, J. Radiotherapy for Colorectal Cancer: Current Standards and Future Perspectives. *Visc. Med.* **32**, 172–177 (2016) doi:10.1159/000446486.

76. Godhi, S., Godhi, A., Bhat, R. & Saluja, S. Colorectal Cancer: Postoperative Follow-up and Surveillance. *Indian J. Surg.* **79**, 234–237 (2017) doi:10.1007/s12262-017-1610-6.

77. Berian, J. R. *et al.* A systematic review of patient perspectives on surveillance after colorectal cancer treatment. *J. Cancer Surviv. Res. Pract.* **11**, 542–552 (2017) doi:10.1007/s11764-017-0623-2.

78. Vera, R. *et al.* Recommendations for follow-up of colorectal cancer survivors. *Clin. Transl. Oncol. Off. Publ. Fed. Span. Oncol. Soc. Natl. Cancer Inst. Mex.* **21**, 1302–1311 (2019) doi:10.1007/s12094-019-02059-1.

79. van Der Stok, E. P., Spaander, M. C. W., Grünhagen, D. J., Verhoef, C. & Kuipers, E. J. Surveillance after curative treatment for colorectal cancer. *Nat. Rev. Oncol.* **14**, 297–315 (2017) doi:10.1038/nrclinonc.2016.199.

80. Bouvier, A. M. *et al.* Incidence and patterns of late recurrences in colon cancer patients. *Int. J. Cancer* **137**, 2133–2138 (2015) doi: 10.1002/ijc.29578.

81. Broadbridge, V. T. *et al.* Do metastatic colorectal cancer patients who present with late relapse after curative surgery have a better survival? *Br. J. Cancer* **109**, 1338 (2013) doi: 10.1038/bjc.2013.388.

82. Cottet, V. *et al.* Incidence and patterns of late recurrences in rectal cancer patients. *Ann. Surg. Oncol.* **22**, 520–527 (2015) doi:10.1245/s10434-014-3990-1.

83. Meyerhardt, J. A. *et al.* Follow-up care, surveillance protocol, and secondary prevention measures for survivors of colorectal cancer: American Society of Clinical Oncology clinical practice guideline endorsement. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **31**, 4465–4470 (2013) doi:10.1200/JCO.2013.50.7442.

84. Steele, S. R. *et al.* Practice Guideline for the Surveillance of Patients After Curative Treatment of Colon and Rectal Cancer. *Dis. Colon Rectum* **58**, 713–725 (2015) doi:10.1097/DCR.0000000000000410.

85. Labianca, R. *et al.* Early colon cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **24 Suppl 6**, vi64-72 (2013) doi:10.1093/annonc/mdt354.

86. Van Cutsem, E., Cervantes, A., Nordlinger, B., Arnold, D. & ESMO Guidelines Working Group. Metastatic colorectal cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **25 Suppl 3**, iii1-9 (2014) doi:10.1093/annonc/mdu260.

87. National Comprehensive Cancer Network. Guidelines for treatment of colorectal cancer. NCCN https://www.nccn.org/professionals/physician_gls/f_guidelines.asp#site (2016).

88. National Institute for Health and Care Excellence. Colorectal cancer: diagnosis and management. NICE http://www.nice.org.uk/Guidance/CG131 (updated 2014).

89. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer* **99**, 260–266 (2002) doi:10.1002/ijc.10332.

90. Fijneman, R. J. A. Genetic predisposition to sporadic cancer: how to handle major effects of minor genes? *Cell. Oncol. Off. J. Int. Soc. Cell. Oncol.* **27**, 281–292 (2005) doi:10.1155/2005/737191.

91. Naccarati, A., Pardini, B., Hemminki, K. & Vodicka, P. Sporadic colorectal cancer and individual susceptibility: a review of the association studies investigating the role of DNA

repair genetic polymorphisms. *Mutat. Res.* **635**, 118–145 (2007) doi:10.1016/j.mrrev.2007.02.001.

92. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39**, 984–988 (2007) doi:10.1038/ng2085.

93. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer* **17**, 692–704 (2017) doi:10.1038/nrc.2017.82.

94. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76–87 (2019) doi:10.1038/s41588-018-0286-6.

95. Liu, G. *et al.* Two novel BRM insertion promoter sequence variants are associated with loss of BRM expression and lung cancer risk. *Oncogene* **30**, 3295–3304 (2011) doi: 10.1038/onc.2011.81.

96. Wong, K. M. *et al.* Two BRM promoter insertion polymorphisms increase the risk of early-stage upper aerodigestive tract cancers. *Cancer Med.* **3**, 426–433 (2014) doi:10.1002/cam4.201.

97. Gao, X. *et al.* Insertion/deletion polymorphisms in the promoter region of BRM contribute to risk of hepatocellular carcinoma in Chinese populations. *PloS One* **8**, e55169 (2013) doi: 10.1371/journal.pone.0055169.

98. Wang, J. R. *et al.* Association of two BRM promoter polymorphisms with head and neck squamous cell carcinoma risk. *Carcinogenesis* **34**, 1012–1017 (2013) doi:10.1093/carcin/bgt008.

99. Lee, M. J. *et al.* Association of two BRM promoter polymorphisms and smoking status with malignant pleural mesothelioma risk and prognosis. *Mol. Carcinog.* **58**, 1960–1973 (2019) doi:10.1002/mc.23088.

100. Al-Tassan, N. A. *et al.* A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. *Sci. Rep.* **5**, 10442 (2015) doi:10.1038/srep10442.

101. Whiffin, N. *et al.* Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. *Hum. Mol. Genet.* **23**, 4729–4737 (2014) doi:10.1093/hmg/ddu177.

102. Houlston, R. S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nat. Genet.* **42**, 973–977 (2010) doi:10.1038/ng.670.

103. Peters, U. *et al.* Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. *Gastroenterology* **144**, 799-807.e24 (2013) doi:10.1053/j.gastro.2012.12.020.

104. Orlando, G. *et al.* Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Hum. Mol. Genet.* **25**, 2349–2359 (2016) doi:10.1093/hmg/ddw087.

105. Schumacher, F. R. *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nat. Commun.* **6**, 7138 (2015) doi:10.1038/ncomms8138.

106. Schmit, S. L. *et al.* Novel Common Genetic Susceptibility Loci for Colorectal Cancer. *J. Natl. Cancer Inst.* **111**, 146–157 (2019) doi:10.1093/jnci/djy099.

107. Schmit, S. L. *et al.* A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. *Carcinogenesis* **35**, 2512–2519 (2014) doi:10.1093/carcin/bgu148.

108. Jia, W.-H. *et al.* Genome-wide association analyses in East Asians identify new susceptibility loci for colorectal cancer. *Nat. Genet.* **45**, 191–196 (2013) doi:10.1038/ng.2505.

109. Zeng, C. *et al.* Identification of Susceptibility Loci and Genes for Colorectal Cancer Risk. *Gastroenterology* **150**, 1633–1645 (2016) doi:10.1053/j.gastro.2016.02.076.

110. Dunlop, M. G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat. Genet.* **44**, 770–776 (2012) doi:10.1038/ng.2293.

111. Cui, R. *et al.* Common variant in 6q26-q27 is associated with distal colon cancer in an Asian population. *Gut* **60**, 799–805 (2011) doi:10.1136/gut.2010.215947.

112. Tomlinson, I. P. M. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**, 623–630 (2008) doi:10.1038/ng.111.

113. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.* **40**, 631–637 (2008) doi: 10.1038/ng.133.

114. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007) doi:10.1038/ng2089.

115. Zhang, B. *et al.* Large-scale genetic study in East Asians identifies six new loci associated with colorectal cancer risk. *Nat. Genet.* **46**, 533–542 (2014) doi: 10.1038/ng.2985.

116. Wang, H. *et al.* Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A. *Nat. Commun.* **5**, 4613 (2014) doi:10.1038/ncomms5613.

117. Wang, M. *et al.* Common genetic variation in ETV6 is associated with colorectal cancer susceptibility. *Nat. Commun.* **7**, 11478 (2016) doi:10.1038/ncomms11478.

118. COGENT Study *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–1435 (2008) doi:10.1038/ng.262.

119. Lemire, M. *et al.* A genome-wide association study for colorectal cancer identifies a risk locus in 14q23.1. *Hum. Genet.* **134**, 1249–1262 (2015) doi:10.1007/s00439-015-1598-6.

120. Tanikawa, C. *et al.* GWAS identifies two novel colorectal cancer loci at 16q24.1 and 20q13.12. *Carcinogenesis* **39**, 652–660 (2018) doi:10.1093/carcin/bgy026.

121. Zhang, B. *et al.* Genome-wide association study identifies a new SMAD7 risk variant associated with colorectal cancer risk in East Asians. *Int. J. Cancer* **135**, 948–955 (2014) doi:10.1002/ijc.28733.

122. Broderick, P. *et al.* A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat. Genet.* **39**, 1315–1317 (2007) doi:10.1038/ng.2007.18.

123. Wang, H. *et al.* Novel colon cancer susceptibility variants identified from a genome-wide association study in African Americans. *Int. J. Cancer* **140**, 2728–2733 (2017) doi:10.1002/ijc.30687.

124. Keum, N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nat. Rev. Gastroenterol. Hepatol.* **16**, 713–732 (2019) doi:10.1038/s41575-019-0189-8.

125. Ajouz, H., Mukherji, D. & Shamseddine, A. Secondary bile acids: an underrecognized cause of colon cancer. *World J. Surg. Oncol.* **12**, 164 (2014) doi:10.1186/1477-7819-12-164.

126. Harris, R. E. *Global epidemiology of cancer*. (Jones & Bartlett Publishers, 2015).

127. Wynder, E. L., Kajitani, T., Ishikawa, S., Dodo, H. & Takano, A. Environmental factors of cancer of the colon and rectum. II. Japanese epidemiological data. *Cancer* **23**, 1210–1220 (1969) doi:10.1002/1097-0142(196905)23:5<1210::aid-cncr2820230530>3.0.co;2-m.

128. Baena, R. & Salinas, P. Diet and colorectal cancer. *Maturitas* **80**, 258–264 (2015) doi:10.1016/j.maturitas.2014.12.017.

129. Song, M. & Chan, A. T. Diet, Gut Microbiota, and Colorectal Cancer Prevention: A Review of Potential Mechanisms and Promising Targets for Future Research. *Curr. Colorectal Cancer Rep.* **13**, 429–439 (2017) doi:10.1007/s11888-017-0389-y.

130. Kune, G. A. The Melbourne Colorectal Cancer Study: reflections on a 30-year experience. *Med. J. Aust.* **193**, 648–652 (2010) doi:10.5694/j.1326-5377.2010.tb04093.x.

131. Mousavi, S. M., Fallah, M., Sundquist, K. & Hemminki, K. Age- and time-dependent changes in cancer incidence among immigrants to Sweden: colorectal, lung, breast and prostate cancers. *Int. J. Cancer* **131**, E122-128 (2012) doi:10.1002/ijc.27334.

132. Haggar, F. A. & Boushey, R. P. Colorectal cancer epidemiology: incidence, mortality, survival, and risk factors. *Clin. Colon Rectal Surg.* **22**, 191–197 (2009) doi: 10.1055/s-0029-1242458.

133. Duthie, S. J. Folic acid deficiency and cancer: mechanisms of DNA instability. *Br. Med. Bull.* **55**, 578–592 (1999) doi:10.1258/0007142991902646.

134. Homann, N., Tillonen, J. & Salaspuro, M. Microbially produced acetaldehyde from ethanol may increase the risk of colon cancer via folate deficiency. *Int. J. Cancer* **86**, 169–173 (2000) doi:10.1002/(sici)1097-0215(20000415)86:2<169::aid-ijc4>3.0.co;2-3.

135. Giovannucci, E. An updated review of the epidemiological evidence that cigarette smoking increases risk of colorectal cancer. *Cancer Epidemiol. Biomark. Prev.* **10**, 725–731 (2001) PMID: 11440957.

136. Lukas, M. Inflammatory bowel disease as a risk factor for colorectal cancer. *Dig. Dis. Basel Switz.* **28**, 619–624 (2010) doi:10.1159/000320276.

137. Compton, C. C. *et al.* Prognostic factors in colorectal cancer: College of American Pathologists consensus statement 1999. *Arch. Pathol. Lab. Med.* **124**, 979–994 (2000) doi: 10.5858/2000-124-0979-PFICC.

138. Zlobec, I. & Lugli, A. Prognostic and predictive factors in colorectal cancer. *Postgrad. Med. J.* **84**, 403–411 (2008) doi:10.1136/jcp.2007.054858.

139. Compton, C. C. & Greene, F. L. The staging of colorectal cancer: 2004 and beyond. *CA. Cancer J. Clin.* **54**, 295–308 (2004) doi:10.3322/canjclin.54.6.295.

140. Weiser, M. R. AJCC 8th Edition: Colorectal Cancer. *Ann. Surg. Oncol.* **25**, 1454–1455 (2018) doi:10.1245/s10434-018-6462-1.

141. Li, J., Wang, Z., Yuan, X., Xu, L. & Tong, J. The prognostic significance of age in operated and non-operated colorectal cancer. *BMC Cancer* **15**, 83 (2015) doi:10.1186/s12885-015-1071-x.

142. Li, M., Li, J. Y., Zhao, A. L. & Gu, J. Colorectal cancer or colon and rectal cancer? Clinicopathological comparison between colonic and rectal carcinomas. *Oncology* **73**, 52–57 (2007) doi:10.1159/000120628.

143. Mehrkhani, F., Nasiri, S., Donboli, K., Meysamie, A. & Hedayat, A. Prognostic factors in survival of colorectal cancer patients after surgery. *Colorectal Dis. Off. J. Assoc. Coloproctology G. B. Irel.* **11**, 157–161 (2009) doi:10.1111/j.1463-1318.2008.01556.x.

144. Mitrovic, B., Schaeffer, D. F., Riddell, R. H. & Kirsch, R. Tumor budding in colorectal carcinoma: time to take notice. *Mod. Pathol.* **25**, 1315–1325 (2012) doi:10.1038/modpathol.2012.94.

145. van Eeghen, E. E., Bakker, S. D., van Bochove, A. & Loffeld, R. J. L. F. Impact of age and comorbidity on survival in colorectal cancer. *J. Gastrointest. Oncol.* **6**, 605–612 (2015) doi:10.3978/j.issn.2078-6891.2015.070.

146. Wang, R., Wang, M.-J. & Ping, J. Clinicopathological Features and Survival Outcomes of Colorectal Cancer in Young Versus Elderly: A Population-Based Cohort Study of SEER 9 Registries Data (1988-2011). *Medicine (Baltimore)* **94**, e1402 (2015) doi:10.1097/MD.0000000000001402.

147. O'Connell, J. B., Maggard, M. A., Livingston, E. H. & Cifford, K. Y. Colorectal cancer in the young. *Am. J. Surg.* **187**, 343–348 (2004) doi: 10.1016/j.amjsurg.2003.12.020.

148. Sharkas, G. F. *et al.* Colorectal Cancer in Jordan: Survival Rate and Its Related Factors. *J. Oncol.* **2017**, 3180762 (2017) doi:10.1155/2017/3180762.

149. Gatalica, Z., Vranic, S., Xiu, J., Swensen, J. & Reddy, S. High microsatellite instability (MSI-H) colorectal carcinoma: a brief review of predictive biomarkers in the era of personalized medicine. *Fam. Cancer* **15**, 405–412 (2016) doi:10.1007/s10689-016-9884-6.

150. Buckowitz, A. *et al.* Microsatellite instability in colorectal cancer is associated with local lymphocyte infiltration and low frequency of distant metastases. *Br. J. Cancer* **92**, 1746 (2005) doi: 10.1038/sj.bjc.6602534.

151. Li, Z.-N., Zhao, L., Yu, L.-F. & Wei, M.-J. BRAF and KRAS mutations in metastatic colorectal cancer: future perspectives for personalized therapy. *Gastroenterol. Rep.* **8**, 192–205 (2020) doi:10.1093/gastro/goaa022.

152. Zarkavelis, G. *et al.* Current and future biomarkers in colorectal cancer. *Ann. Gastroenterol.* **30**, 613–621 (2017) doi:10.20524/aog.2017.0191.

153. Guo, Y.-J. *et al.* ERK/MAPK signalling pathway and tumorigenesis. *Exp. Ther. Med.* **19**, 1997–2007 (2020) doi:10.3892/etm.2020.8454.

154. Santarpia, L., Lippman, S. M. & El-Naggar, A. K. Targeting the MAPK-RAS-RAF signaling pathway in cancer therapy. *Expert Opin. Ther. Targets* **16**, 103–119 (2012) doi:10.1517/14728222.2011.645805.

155. Foltran, L. *et al.* Prognostic role of KRAS, NRAS, BRAF and PIK3CA mutations in advanced colorectal cancer. *Future Oncol. Lond. Engl.* **11**, 629–640 (2015) doi:10.2217/fon.14.279.

156. Kadowaki, S. *et al.* Prognostic value of KRAS and BRAF mutations in curatively resected colorectal cancer. *World J. Gastroenterol.* **21**, 1275–1283 (2015) doi:10.3748/wjg.v21.i4.1275.

157. Modest, D. P. *et al.* Outcome according to KRAS-, NRAS- and BRAF-mutation as well as KRAS mutation variants: pooled analysis of five randomized trials in metastatic colorectal cancer by the AIO colorectal cancer study group. *Ann. Oncol. Off. J. Eur. Soc. Med. Oncol.* **27**, 1746–1753 (2016) doi:10.1093/annonc/mdw261.

158. Ahn, T. S. *et al.* The BRAF mutation is associated with the prognosis in colorectal cancer. *J. Cancer Res. Clin. Oncol.* **140**, 1863–1871 (2014) doi:10.1007/s00432-014-1735-y.

159. Yokota, T. *et al.* BRAF mutation is a powerful prognostic factor in advanced and recurrent colorectal cancer. *Br. J. Cancer* **104**, 856–862 (2011) doi:10.1038/bjc.2011.19.

160. Alves Martins, B. A. *et al.* Biomarkers in Colorectal Cancer: The Role of Translational Proteomics Research. *Front. Oncol.* **9**, 1284 (2019) doi:10.3389/fonc.2019.01284.

161. Nakagoe, T. *et al.* Prognostic value of carcinoembryonic antigen (CEA) in tumor tissue of patients with colorectal cancer. *Anticancer Res.* **21**, 3031–3036 (2001) PMID: 11712806.

162. Coghlin, C. & Murray, G. I. Biomarkers of colorectal cancer: recent advances and future challenges. *Proteomics Clin. Appl.* **9**, 64–71 (2015) doi:10.1002/prca.201400082.

163. To, K. K., Tong, C. W., Wu, M. & Cho, W. C. MicroRNAs in the prognosis and therapy of colorectal cancer: From bench to bedside. *World J. Gastroenterol.* **24**, 2949–2973 (2018) doi:10.3748/wjg.v24.i27.2949.

164. Afzal, S. *et al.* The association of polymorphisms in 5-fluorouracil metabolism genes with outcome in adjuvant treatment of colorectal cancer. *Pharmacogenomics* **12**, 1257–1267 (2011) doi:10.2217/pgs.11.83.

165. Curtin, K. *et al.* Thymidylate synthase polymorphisms and colon cancer: associations with tumor stage, tumor characteristics and survival. *Int. J. Cancer* **120**, 2226–2232 (2007) doi:10.1002/ijc.22603.

166. Ose, J. *et al.* Pathway analysis of genetic variants in folate-mediated one-carbon metabolism-related genes and survival in a prospectively followed cohort of colorectal cancer patients. *Cancer Med.* (2018) doi:10.1002/cam4.1407.

167. Dai, J. *et al.* GWAS-identified colorectal cancer susceptibility loci associated with clinical outcomes. *Carcinogenesis* **33**, 1327–1331 (2012) doi:10.1093/carcin/bgs147.

168. He, Y. *et al.* Effects of common genetic variants associated with colorectal cancer risk on survival outcomes after diagnosis: A large population-based cohort study. *Int. J. Cancer* **145**, 2427–2432 (2019) doi:10.1002/ijc.32550.

169. Hu, Y. *et al.* Colorectal cancer susceptibility loci as predictive markers of rectal cancer prognosis after surgery. *Genes. Chromosomes Cancer* **57**, 140–149 (2018) doi:10.1002/gcc.22512.

170. Kang, B. W. *et al.* Association between GWAS-identified genetic variations and disease prognosis for patients with colorectal cancer. *PLoS One* **10**, e0119649 (2015) doi:10.1371/journal.pone.0119649.

171. Smith, C. G. *et al.* Analyses of 7,635 patients with colorectal cancer using independent training and validation cohorts show that rs9929218 in CDH1 is a prognostic marker of survival. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **21**, 3453–3461 (2015) doi:10.1158/1078-0432.CCR-14-3136.

172. Song, N. *et al.* Colorectal cancer susceptibility loci and influence on survival. *Genes. Chromosomes Cancer* **57**, 630–637 (2018) doi:10.1002/gcc.22674.

173. Summers, M. G. *et al.* Comprehensive analysis of colorectal cancer-risk loci and survival outcome: A prognostic role for CDH1 variants. *Eur. J. Cancer* **124**, 56–63 (2020) doi:10.1016/j.ejca.2019.09.024.

174. Xing, J. *et al.* GWAS-identified colorectal cancer susceptibility locus associates with disease prognosis. *Eur. J. Cancer Oxf. Engl. 1990* **47**, 1699–1707 (2011) doi:10.1016/j.ejca.2011.02.004.

175. Scherer, D. *et al.* Polymorphisms in the Angiogenesis-Related Genes EFNB2, MMP2 and JAG1 Are Associated with Survival of Colorectal Cancer Patients. *Int. J. Mol. Sci.* **21**, (2020) doi:10.3390/ijms21155395.

176. Xu, W. *et al.* A genome wide association study on Newfoundland colorectal cancer patients' survival outcomes. *Biomark. Res.* **3**, 6 (2015) doi: 10.1186/s40364-015-0031-6.

177. Penney, M. E., Parfrey, P. S., Savas, S. & Yilmaz, Y. E. A genome-wide association study identifies single nucleotide polymorphisms associated with time-to-metastasis in colorectal cancer. *BMC Cancer* **19**, 133 (2019) doi:10.1186/s12885-019-5346-5.

178. Pander, J. *et al.* Genome wide association study for predictors of progression free survival in patients on capecitabine, oxaliplatin, bevacizumab and cetuximab in first-line therapy of metastatic colorectal cancer. *PLoS One* **10**, e0131091 (2015) doi:10.1371/journal.pone.0131091.

179. Phipps, A. I. *et al.* Common genetic variation and survival after colorectal cancer diagnosis: a genome-wide analysis. *Carcinogenesis* **37**, 87–95 (2016) doi:10.1093/carcin/bgv161.

180. Penney, K. L. *et al.* Genetic variant associated with survival of patients with stage II-III colon cancer. *Clin. Gastroenterol. Hepatol. Off. Clin. Pract. J. Am. Gastroenterol. Assoc.* **18**, 2717-2723.e3 (2020) doi:10.1016/j.cgh.2019.11.046.

181. Innocenti, F. *et al.* Genomic analysis of germline variation associated with survival of colorectal cancer patients treated with chemotherapy plus biologics in CALGB/SWOG 80405 (Alliance). *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* (2020) doi:10.1158/1078-0432.CCR-20-2021.

182. Carroll, O. U., Morris, T. P. & Keogh, R. H. How are missing data in covariates handled in observational time-to-event studies in oncology? A systematic review. *BMC Med. Res. Methodol.* **20**, 134 (2020) doi:10.1186/s12874-020-01018-7.

183. Jachno, K., Heritier, S. & Wolfe, R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised controlled trials? A review of current practice. *BMC Med. Res. Methodol.* **19**, 103 (2019) doi:10.1186/s12874-019-0749-1.

184. Altman, D. G., De Stavola, B. L., Love, S. B. & Stepniewska, K. A. Review of survival analyses published in cancer journals. *Br. J. Cancer* **72**, 511 (1995) doi: 10.1038/bjc.1995.364.

185. Bellera, C. A. *et al.* Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Med. Res. Methodol.* **10**, 20 (2010) doi: 10.1186/1471-2288-10-20.

186. Dekker, F. W., de Mutsert, R., van Dijk, P. C., Zoccali, C. & Jager, K. J. Survival analysis: time-dependent effects and time-varying risk factors. *Kidney Int.* **74**, 994–997 (2008) doi:10.1038/ki.2008.328.

187. He, K., Yang, Y., Li, Y., Zhu, J. & Li, Y. Modeling Time-Varying Effects With Large-Scale Survival Data: An Efficient Quasi-Newton Approach. *J. Comput. Graph. Stat.* **26**, 635–645 (2017) doi:10.1080/10618600.2016.1237364.

188. Zhang, Z., Reinikainen, J., Adeleke, K. A., Pieterse, M. E. & Groothuis-Oudshoorn, C. G. M. Time-varying covariates and coefficients in Cox regression models. *Ann. Transl. Med.* **6**, 121 (2018) doi:10.21037/atm.2018.02.12.

189. Yu, Y. *et al.* The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. *BMC Med.* **17**, 150 (2019) doi:10.1186/s12916-019-1379-5.

190. Roncucci, L. *et al.* Survival for colon and rectal cancer in a population-based cancer registry. *Eur. J. Cancer* **32A**, 295–302 (1996) doi: 10.1016/0959-8049(95)00532-3.

191. Bolard, P., Quantin, C., Esteve, J., Faivre, J. & Abrahamowicz, M. Modelling time-dependent hazard ratios in relative survival: application to colon cancer. *J. Clin. Epidemiol.* **54**, 986–996 (2001) doi: 10.1016/s0895-4356(01)00363-8.

192. Giorgi, R. *et al.* A relative survival regression model using B-spline functions to model non-proportional hazards. *Stat. Med.* **22**, 2767–2784 (2003) doi: 10.1002/sim.1484.

193. Quantin, C. *et al.* Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *Am. J. Epidemiol.* **150**, 1188–1200 (1999) doi: 10.1093/oxfordjournals.aje.a009945.

194. Zahl, P. Regression analysis with multiplicative and time-varying additive regression coefficients with examples from breast and colon cancer. *Stat. Med.* **22**, 1113–1127 (2003) doi: 10.1002/sim.971.

195. Liu, M. *et al.* Marriage is a dependent risk factor for mortality of colon adenocarcinoma without a time-varying effect. *Oncotarget* **8**, 20056–20066 (2017) doi:10.18632/oncotarget.15378.

196. Pavelitz, T. *et al.* MRE11-deficiency associated with improved long-term disease free survival and overall survival in a subset of stage III colon cancer patients in randomized CALGB 89803 trial. *PLoS One* **9**, e108483 (2014) doi: 10.1371/journal.pone.0108483.

197. Gooiker, G. A. *et al.* Risk factors for excess mortality in the first year after curative surgery for colorectal cancer. *Ann. Surg. Oncol.* **19**, 2428–2434 (2012) doi:10.1245/s10434-012-2294-6.

198. Artinyan, A. *et al.* Infectious postoperative complications decrease long-term survival in patients undergoing curative surgery for colorectal cancer: a study of 12,075 patients. *Ann. Surg.* **261**, 497–505 (2015) doi:10.1097/SLA.0000000000000854.

199. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B Methodol.* **34**, 187–220 (1972) doi:10.1111/j.2517-6161.1972.tb00899.x.

200. Cox, D. R. & Oakes, D. *Analysis of survival data*. (Chapman and Hall, 1984) doi: 10.1201/9781315137438.

201. Hernán, M. A. The hazards of hazard ratios. *Epidemiol. Camb. Mass* **21**, 13–15 (2010) doi:10.1097/EDE.0b013e3181c1ea43.

202. Xue, Y. & Schifano, E. D. Diagnostics for the Cox model. *Commun. Stat. Appl. Methods* **24**, 583–604 (2017) doi:10.29220/CSAM.2017.24.6.583.

203. Hess, K. R. Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Stat. Med.* **14**, 1707–1723 (1995) doi:10.1002/sim.4780141510.

204. Grambsch, P. M. & Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526 (1994) doi: 10.1093/biomet/81.3.515.

205. Therneau, T. M. A package for survival analysis in S. version 2.38. (2015) Available at: https://CRAN.R-project.org/package=survival (accessed on Nov 23, 2021).

206. R Development Core Team. R: A language and environment for statistical computing. (2013) Available at: https://www.r-project.org/ (accessed on Nov 23, 2021).

207. Ng'andu, N. H. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Stat. Med.* **16**, 611–626 (1997) doi:10.1002/(sici)1097-0258(19970330)16:6<611::aid-sim437>3.0.co;2-t.

208. Schemper, M. Cox Analysis of Survival Data with Non-Proportional Hazard Functions. *The Statistician* **41**, 455 (1992) doi:10.2307/2349009.

209. Harrell, F. E. Cox Proportional Hazards Regression Model. in *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* (ed. Harrell, Jr., Frank E.) 475–519 (Springer International Publishing, 2015). doi:10.1007/978-3-319-19425-7_20.

210. Klein, J. P. & Moeschberger, M. L. Refinements of the semiparametric proportional hazards model. in *Survival analysis: techniques for censored and truncated data* 295–328 (Springer, 2003) doi: 10.1007/0-387-21645-6_9.

211. Kleinbaum, D. G. & Klein, M. The Stratified Cox Procedure. in *Survival Analysis: A Self-Learning Text* 201–240 (Springer, 2012). doi:10.1007/978-1-4419-6646-9_5.

212. Matthews, D. E. & Farewell, V. T. On testing for a constant hazard against a change-point alternative. *Biometrics* **38**, 463–468 (1982) doi: 10.2307/2530460.

213. Buchholz, A. & Sauerbrei, W. Comparison of procedures to assess non-linear and time-varying effects in multivariable models for survival data. *Biom. J. Biom. Z.* **53**, 308–331 (2011) doi:10.1002/bimj.201000159.

214. Thomas, L. & Reyes, E. M. Tutorial: Survival Estimation for Cox Regression Models with Time-Varying Coefficients Using SAS and R. *J. Stat. Softw.* **61**, (2014) doi:10.18637/jss.v061.c01.

215. Ahmed, F. E., Vos, P. W. & Holbert, D. Modeling survival in colon cancer: a methodological review. *Mol. Cancer* **6**, 15 (2007) doi:10.1186/1476-4598-6-15.

216. Gray, R. J. Flexible Methods for Analyzing Survival Data Using Splines, with Applications to Breast Cancer Prognosis. *J. Am. Stat. Assoc.* **87**, 942–951 (1992) doi:10.1080/01621459.1992.10476248.

217. Klein, J. P. & Moeschberger, M. L. Inference for Parametric Regression Models. in *Survival analysis: techniques for censored and truncated data* 393–423 (Springer, 2003). doi:10.1007/0-387-21645-6_12.

218. Kleinbaum, D. G. & Klein, M. *Survival analysis: a self-learning text*. (Springer, 2012) doi: 10.1007/978-1-4419-6646-9.

219. James, I. Accelerated Failure-time Models. in *Wiley StatsRef: Statistics Reference Online* (eds. Balakrishnan, N. et al.) stat06002 (John Wiley & Sons, Ltd, 2014). doi:10.1002/9781118445112.stat06002.

220. Orbe, J., Ferreira, E. & Núñez-Antón, V. Comparing proportional hazards and accelerated failure time models for survival analysis: COMPARING MODELS FOR SURVIVAL ANALYSIS. *Stat. Med.* **21**, 3493–3510 (2002) doi:10.1002/sim.1251.

221. Hosmer, D. W., Lemeshow, S. & May, S. Parametric Regression Models. in *Applied Survival Analysis: regression modeling of time-to-event data* 244–285 (John Wiley & Sons, Inc., 2011). doi:10.1002/9780470258019.ch8.

222. Balakrishnan, N., Chimitova, E., Galanova, N. & Vedernikova, M. Testing Goodness of Fit of Parametric AFT and PH Models with Residuals. *Commun. Stat. - Simul. Comput.* **42**, 1352–1367 (2013) doi:10.1080/03610918.2012.659824.

223. Başar, E. Aalen's Additive, Cox Proportional Hazards and the Cox-Aalen Model: Application to Kidney Transplant Data. *Sains Malays.* **46**, 469–476 (2017) doi:10.17576/jsm-2017-4603-15.

224. Xie, X., Strickler, H. D. & Xue, X. Additive hazard regression models: an application to the natural history of human papillomavirus. *Comput. Math. Methods Med.* **2013**, 796270 (2013) doi:10.1155/2013/796270.

225. Aalen, O. O. A Model for Nonparametric Regression Analysis of Counting Processes. in *Lecture Notes in Statistics-2: Mathematical Statistics and Probability Theory* (eds. Klonecki, W., Kozek, A. & Rosiński, J.) 1–25 (Springer New York, 1980) doi: 10.1007/978-1-4615-7397-5_1.

226. Klein, J. P. & Moeschberger, M. L. Additive Hazard Regression Models. in *Survival Analysis: Techniques for Censored and Truncated Data* 329–352 (Springer, 2003). doi:10.1007/0-387-21645-6_10.

227. Scheike, T. H. & Zhang, M.-J. An Additive-Multiplicative Cox-Aalen Regression Model. *Scand. J. Stat.* **29**, 75–88 (2002) doi:10.1111/1467-9469.00065.

228. Amico, M. & Van Keilegom, I. Cure Models in Survival Analysis. *Annu. Rev. Stat. Its Appl.* **5**, 311–342 (2018) doi:10.1146/annurev-statistics-031017-100101.

229. Othus, M., Barlogie, B., LeBlanc, M. L. & Crowley, J. J. Cure Models as a Useful Statistical Tool for Analyzing Survival. *Clin. Cancer Res.* **18**, 3731–3736 (2012) doi:10.1158/1078-0432.CCR-11-2859.

230. Reich, D. E. *et al.* Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**, 135–142 (2002) doi:10.1038/ng947.

231. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003) doi:10.1038/nature02168.

232. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005) doi:10.1038/nature04226.

233. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007) doi:10.1038/nature06258.

234. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010) doi:10.1038/nature09298.

235. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015) doi:10.1038/nature15393.

236. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020) doi:10.1038/s41586-020-2308-7.

237. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nat. Rev. Genet.* **17**, 93–108 (2016) doi:10.1038/nrg.2015.17.

238. Zhang, F. & Lupski, J. R. Non-coding genetic variants in human disease. *Hum. Mol. Genet.* **24**, R102–R110 (2015) doi:10.1093/hmg/ddv259.

239. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019) doi:10.1093/nar/gky1120.

240. Collins, R. L. *et al.* A structural variation reference for medical and population genetics. *Nature* **581**, 444–451 (2020) doi:10.1038/s41586-020-2287-8.

241. Komar, A. A. Silent SNPs: impact on gene function and phenotype. *Pharmacogenomics* **8**, 1075–1080 (2007) doi:10.2217/14622416.8.8.1075.

242. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015) doi:10.1038/nrg3891.

243. Zhang, Y., Qian, J., Gu, C. & Yang, Y. Alternative splicing and cancer: a systematic review. *Signal Transduct. Target. Ther.* **6**, 78 (2021) doi:10.1038/s41392-021-00486-7.

244. Cai, Y., Yu, X., Hu, S. & Yu, J. A brief review on the mechanisms of miRNA regulation. *Genomics Proteomics Bioinformatics* **7**, 147–154 (2009) doi:10.1016/S1672-0229(08)60044-3.

245. Westra, H.-J. & Franke, L. From genome to function by studying eQTLs. *Biochim. Biophys. Acta* **1842**, 1896–1902 (2014) doi:10.1016/j.bbadis.2014.04.024.

246. Umans, B. D., Battle, A. & Gilad, Y. Where Are the Disease-Associated eQTLs? *Trends Genet. TIG* **37**, 109–124 (2021) doi:10.1016/j.tig.2020.08.009.

247. Kaplow, I. M. *et al.* A pooling-based approach to mapping genetic variants associated with DNA methylation. *Genome Res.* **25**, 907–917 (2015) doi:10.1101/gr.183749.114.

248. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* **19**, 48–54 (2016) doi:10.1038/nn.4182.

249. Gallagher, M. D. & Chen-Plotkin, A. S. The Post-GWAS Era: From Association to Function. *Am. J. Hum. Genet.* **102**, 717–730 (2018) doi:10.1016/j.ajhg.2018.04.002.

250. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012) doi:10.1038/nature10808.

251. Ding, Z. *et al.* Quantitative genetics of CTCF binding reveal local sequence effects and different modes of X-chromosome association. *PLoS Genet.* **10**, e1004798 (2014) doi:10.1371/journal.pgen.1004798.

252. Tehranchi, A. K. *et al.* Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell* **165**, 730–741 (2016) doi:10.1016/j.cell.2016.03.041.

253. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012) doi:10.1126/science.1222794.

254. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010) doi:10.1038/nature09266.

255. Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat. Genet.* **41**, 882–884 (2009) doi:10.1038/ng.403.

256. Hu, Z. *et al.* Genetic variants of miRNA sequences and non-small cell lung cancer survival. *J. Clin. Invest.* **118**, 2600–2608 (2008) doi:10.1172/JCI34934.

257. Shen, J. *et al.* Genetic polymorphisms in telomere pathway genes, telomere length, and breast cancer survival. *Breast Cancer Res. Treat.* **134**, 393–400 (2012) doi:10.1007/s10549-012-2058-9.

258. Jain, L. *et al.* The role of vascular endothelial growth factor SNPs as predictive and prognostic markers for major solid tumors. *Mol. Cancer Ther.* **8**, 2496–2508 (2009) doi:10.1158/1535-7163.MCT-09-0302.

259. Eng, L. *et al.* Vascular endothelial growth factor pathway polymorphisms as prognostic and pharmacogenetic factors in cancer: a systematic review and meta-analysis. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **18**, 4526–4537 (2012) doi:10.1158/1078-0432.CCR-12-1315.

260. Breen, E. C. VEGF in biological control. *J. Cell. Biochem.* **102**, 1358–1367 (2007) doi:10.1002/jcb.21579.

261. Huang, E. Y. *et al.* A common regulatory variant in SLC35B4 influences the recurrence and survival of prostate cancer. *J. Cell. Mol. Med.* **22**, 3661–3670 (2018) doi:10.1111/jcmm.13649.

262. Negandhi, A. A. *et al.* MTHFR Glu429Ala and ERCC5 His46His polymorphisms are associated with prognosis in colorectal cancer patients: analysis of two independent cohorts from Newfoundland. *PLoS One* **8**, e61469 (2013) doi: 10.1371/journal.pone.0061469.

263. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986-992 (2014) doi:10.1093/nar/gkt958.

264. Carvalho, C. M. B. & Lupski, J. R. Mechanisms underlying structural variant formation in genomic disorders. *Nat. Rev. Genet.* **17**, 224–238 (2016) doi:10.1038/nrg.2015.25.

265. Zhang, F., Gu, W., Hurles, M. E. & Lupski, J. R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009) doi:10.1146/annurev.genom.9.081307.164217.

266. Li, X., Liu, Y., Lu, J. & Zhao, M. Integrative analysis to identify oncogenic gene expression changes associated with copy number variations of enhancer in ovarian cancer. *Oncotarget* **8**, 91558–91567 (2017) doi:10.18632/oncotarget.21227.

267. Klopocki, E. & Mundlos, S. Copy-number variations, noncoding sequences, and human phenotypes. *Annu. Rev. Genomics Hum. Genet.* **12**, 53–72 (2011) doi:10.1146/annurev-genom-082410-101404.

268. Gaszner, M. & Felsenfeld, G. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* **7**, 703–713 (2006) doi:10.1038/nrg1925.

269. Rigau, M., Juan, D., Valencia, A. & Rico, D. Intronic CNVs and gene expression variation in human populations. *PLoS Genet.* **15**, e1007902 (2019) doi:10.1371/journal.pgen.1007902.

270. Li, H. & Anderson, S. K. Association of TNFRSF1B Promoter Polymorphisms with Human Disease: Further Studies Examining T-Regulatory Cells Are Required. *Front. Immunol.* **9**, 443 (2018) doi:10.3389/fimmu.2018.00443.

271. Meyer, N. & Penn, L. Z. Reflecting on 25 years with MYC. *Nat. Rev. Cancer* **8**, 976–990 (2008) doi:10.1038/nrc2231.

272. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012) doi:10.1016/j.cell.2012.08.026.

273. Nie, Z. *et al.* c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. *Cell* **151**, 68–79 (2012) doi:10.1016/j.cell.2012.08.033.

274. Latchman, D. S. Transcription factors: an overview. *Int. J. Exp. Pathol.* **74**, 417–422 (1993) PMID: 8217775.

275. Freund, C., Horsford, D. J. & McInnes, R. R. Transcription factor genes and the developing eye: a genetic perspective. *Hum. Mol. Genet.* **5**, 1471–1488 (1996) doi:10.1093/hmg/5.supplement_1.1471.

276. Weischenfeldt, J., Symmons, O., Spitz, F. & Korbel, J. O. Phenotypic impact of genomic structural variation: insights from and for human disease. *Nat. Rev. Genet.* **14**, 125–138 (2013) doi: 10.1038/nrg3373.

277. Spielmann, M., Lupiáñez, D. G. & Mundlos, S. Structural variation in the 3D genome. *Nat. Rev. Genet.* **19**, 453–467 (2018) doi:10.1038/s41576-018-0007-0.

278. Wilson, B. G. & Roberts, C. W. M. SWI/SNF nucleosome remodellers and cancer. *Nat. Rev. Cancer* **11**, 481–492 (2011) doi:10.1038/nrc3068.

279. Savas, S. & Skardasi, G. The SWI/SNF complex subunit genes: Their functions, variations, and links to risk and survival outcomes in human cancers. *Crit. Rev. Oncol. Hematol.* **123**, 114–131 (2018) doi:10.1016/j.critrevonc.2018.01.009.

280. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010) doi: 10.1146/annurev-med-100708-204735.

281. Bayés, M., Magano, L. F., Rivera, N., Flores, R. & Pérez Jurado, L. A. Mutational mechanisms of Williams-Beuren syndrome deletions. *Am. J. Hum. Genet.* **73**, 131–151 (2003) doi:10.1086/376565.

282. Cao, A. & Galanello, R. Beta-thalassemia. *Genet. Med.* **12**, 61–76 (2010) doi:10.1097/GIM.0b013e3181cd68ed.

283. Harteveld, C. L. & Higgs, D. R. α-thalassaemia. *Orphanet J. Rare Dis.* **5**, 13 (2010) doi:10.1186/1750-1172-5-13.

284. Muncie, H. L. & Campbell, J. Alpha and beta thalassemia. *Am. Fam. Physician* **80**, 339–344 (2009) PMID: 19678601.

285. Nakatochi, M., Kushima, I. & Ozaki, N. Implications of germline copy-number variations in psychiatric disorders: review of large-scale genetic studies. *J. Hum. Genet.* **66**, 25–37 (2021) doi:10.1038/s10038-020-00838-1.

286. Fellermann, K. *et al.* A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am. J. Hum. Genet.* **79**, 439–448 (2006) doi:10.1086/505915.

287. McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat. Genet.* **40**, 1107–1112 (2008) doi:10.1038/ng.215.

288. Chen, Y. & Chen, C. DNA copy number variation and loss of heterozygosity in relation to recurrence of and survival from head and neck squamous cell carcinoma: a review. *Head Neck* **30**, 1361–1383 (2008) doi: 10.1002/hed.20861.

289. Reis, G. F. *et al.* CDKN2A loss is associated with shortened overall survival in lower-grade (World Health Organization Grades II-III) astrocytomas. *J. Neuropathol. Exp. Neurol.* **74**, 442–452 (2015) doi:10.1097/NEN.0000000000000188.

290. Bortolotto, S. *et al.* CDKN2A/p16 inactivation in the prognosis of oligodendrogliomas. *Int. J. Cancer* **88**, 554–557 (2000) doi:10.1002/1097-0215(20001115)88:4<554::aid-ijc6>3.0.co;2-q.

291. Cairncross, J. G. *et al.* Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendrogliomas. *J. Natl. Cancer Inst.* **90**, 1473–1479 (1998) doi:10.1093/jnci/90.19.1473.

292. Bi, H. *et al.* Copy number variation of E3 ubiquitin ligase genes in peripheral blood leukocyte and colorectal cancer. *Sci. Rep.* **6**, 29869 (2016) doi:10.1038/srep29869.

293. Garziera, M. *et al.* HLA-G 3'UTR polymorphisms impact the prognosis of stage II-III CRC patients in fluoropyrimidine-based treatment. *PLoS One* **10**, e0144000 (2015) doi:10.1371/journal.pone.0144000.

294. Lee, K. S. *et al.* c-MYC Copy-Number Gain Is an Independent Prognostic Factor in Patients with Colorectal Cancer. *PloS One* **10**, e0139727 (2015) doi:10.1371/journal.pone.0139727.

295. Jeong, H. M. *et al.* Targeted exome sequencing of Korean triple-negative breast cancer reveals homozygous deletions associated with poor prognosis of adjuvant chemotherapy-treated patients. *Oncotarget* **8**, 61538–61550 (2017) doi:10.18632/oncotarget.18618.

296. Yin, J. *et al.* Copy-number variation of MCL1 predicts overall survival of non-small-cell lung cancer in a Southern Chinese population. *Cancer Med.* **5**, 2171–2179 (2016) doi:10.1002/cam4.774.

297. Kang, M. J. *et al.* 22q11-q13 as a hot spot for prediction of disease-free survival in bile duct cancer: integrative analysis of copy number variations. *Cancer Genet.* **207**, 57–69 (2014) doi:10.1016/j.cancergen.2014.02.003.

298. Gu, X. *et al.* Copy number variation: A prognostic marker for young patients with squamous cell carcinoma of the oral tongue. *J. Oral Pathol. Med. Off. Publ. Int. Assoc. Oral Pathol. Am. Acad. Oral Pathol.* **48**, 24–30 (2019) doi:10.1111/jop.12792.

299. Yu, Y. *et al.* Genome-wide copy number variation analysis identified ANO1 as a novel oncogene and prognostic biomarker in esophageal squamous cell cancer. *Carcinogenesis* **40**, 1198–1208 (2019) doi:10.1093/carcin/bgz077.

300. Bai, J. *et al.* Whole genome sequencing of skull-base chordoma reveals genomic alterations associated with recurrence and chordoma-specific survival. *Nat. Commun.* **12**, 757 (2021) doi:10.1038/s41467-021-21026-5.

301. Stover, D. G. *et al.* Association of Cell-Free DNA Tumor Fraction and Somatic Copy Number Alterations With Survival in Metastatic Triple-Negative Breast Cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **36**, 543–553 (2018) doi:10.1200/JCO.2017.76.0033.

302. Zhang, L. *et al.* Genomic analysis of primary and recurrent gliomas reveals clinical outcome related molecular features. *Sci. Rep.* **9**, 16058 (2019) doi:10.1038/s41598-019-52515-9.

303. Muñoz-Hidalgo, L. *et al.* Somatic copy number alterations are associated with EGFR amplification and shortened survival in patients with primary glioblastoma. *Neoplasia N. Y. N* **22**, 10–21 (2020) doi:10.1016/j.neo.2019.09.001.

304. Rosenberg, S. *et al.* Machine Learning for Better Prognostic Stratification and Driver Gene Identification Using Somatic Copy Number Variations in Anaplastic Oligodendroglioma. *The Oncologist* **23**, 1500–1510 (2018) doi:10.1634/theoncologist.2017-0495.

305. Chen, S. *et al.* Prognostic Value of Germline Copy Number Variants and Environmental Exposures in Non-small Cell Lung Cancer. *Front. Genet.* **12**, 681857 (2021) doi:10.3389/fgene.2021.681857.

306. Sapkota, Y. *et al.* Germline DNA copy number aberrations identified as potential prognostic factors for breast cancer recurrence. *PloS One* **8**, e53850 (2013) doi:10.1371/journal.pone.0053850.

307. Kumaran, M. *et al.* Germline copy number variations are associated with breast cancer risk and prognosis. *Sci. Rep.* **7**, 14621 (2017) doi:10.1038/s41598-017-14799-7.

308. Wain, L. V., Armour, J. A. L. & Tobin, M. D. Genomic copy number variation, human health, and disease. *Lancet Lond. Engl.* **374**, 340–350 (2009) doi:10.1016/S0140-6736(09)60249-X.

309. Guo, H. *et al.* Genome-wide copy number variation analysis in a Chinese autism spectrum disorder cohort. *Sci. Rep.* **7**, 44155 (2017) doi:10.1038/srep44155.

310. Li, D. *et al.* Genome-Wide Association Study of Copy Number Variations (CNVs) with Opioid Dependence. *Neuropsychopharmacology* **40**, 1016–1026 (2015) doi:10.1038/npp.2014.290.

311. Li, Y. R. *et al.* Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nat. Commun.* **11**, 255 (2020) doi:10.1038/s41467-019-13624-1.

312. Say, Y.-H. The association of insertions/deletions (INDELs) and variable number tandem repeats (VNTRs) with obesity and its related traits and complications. *J. Physiol. Anthropol.* **36**, 25 (2017) doi:10.1186/s40101-017-0142-x.

313. Canadian Cancer Society's Advisory Committee on Cancer Statistics. *Canadian Cancer Statistics 2017* (2017) Available at: https://publications.gc.ca/collections/collection_2017/statcan/CS2-37-2017-eng.pdf (accessed on Nov 23, 2021).

314. Sankaranarayanan, R. *et al.* Cancer survival in Africa, Asia, and Central America: a population-based study. *Lancet Oncol.* **11**, 165–173 (2010) doi:10.1016/S1470-2045(09)70335-3.

315. Jiang, K. *et al.* Genome-wide association study identifies two new susceptibility loci for colorectal cancer at 5q23.3 and 17q12 in Han Chinese. *Oncotarget* **6**, 40327–40336 (2015) doi:10.18632/oncotarget.5530.

316. Muchardt, C. & Yaniv, M. A human homologue of Saccharomyces cerevisiae SNF2/SWI2 and Drosophila brm genes potentiates transcriptional activation by the glucocorticoid receptor. *EMBO J.* **12**, 4279–4290 (1993) doi: 10.1002/j.1460-2075.1993.tb06112.x.

317. Peterson, C. L. & Workman, J. L. Promoter targeting and chromatin remodeling by the SWI/SNF complex. *Curr. Opin. Genet. Dev.* **10**, 187–192 (2000) doi: 10.1016/s0959-437x(00)00068-x.

318. Reisman, D., Glaros, S. & Thompson, E. A. The SWI/SNF complex and cancer. *Oncogene* **28**, 1653–1668 (2009) doi:10.1038/onc.2009.4.

319. Glaros, S. *et al.* The reversible epigenetic silencing of BRM: implications for clinical targeted therapy. *Oncogene* **26**, 7058–7066 (2007) doi:10.1038/sj.onc.1210514.

320. Herpel, E. *et al.* SMARCA4 and SMARCA2 deficiency in non-small cell lung cancer: immunohistochemical survey of 316 consecutive specimens. *Ann. Diagn. Pathol.* **26**, 47–51 (2017) doi:10.1016/j.anndiagpath.2016.10.006.

321. Matsubara, D. *et al.* Lung cancer with loss of BRG1/BRM, shows epithelial mesenchymal transition phenotype and distinct histologic and genetic features. *Cancer Sci.* **104**, 266–273 (2013) doi:10.1111/cas.12065.

322. Reisman, D. N., Sciarrotta, J., Wang, W., Funkhouser, W. K. & Weissman, B. E. Loss of BRG1/BRM in human lung cancer cell lines and primary lung cancers: correlation with poor prognosis. *Cancer Res.* **63**, 560–566 (2003) PMID: 12566296.

323. Kahali, B. *et al.* The silencing of the SWI/SNF subunit and anticancer gene BRM in Rhabdoid tumors. *Oncotarget* **5**, 3316–3332 (2014) doi:10.18632/oncotarget.1945.

324. Korpanty, G. J. *et al.* Association of BRM promoter polymorphisms and esophageal adenocarcinoma outcome. *Oncotarget* **8**, 28093–28100 (2017) doi:10.18632/oncotarget.15890.

325. Liu, G. *et al.* BRM Promoter Polymorphisms and Survival of Advanced Non-Small Cell Lung Cancer Patients in the Princess Margaret Cohort and CCTG BR.24 Trial. *Clin.*

Cancer Res. Off. J. Am. Assoc. Cancer Res. **23**, 2460–2470 (2017) doi:10.1158/1078-0432.CCR-16-1640.

326. Pasic, I. *et al.* Two BRM promoter polymorphisms predict poor survival in patients with hepatocellular carcinoma. *Mol. Carcinog.* **57**, 106–113 (2018) doi:10.1002/mc.22736.
327. Segedi, M. *et al.* BRM polymorphisms, pancreatic cancer risk and survival. *Int. J. Cancer* **139**, 2474–2481 (2016) doi:10.1002/ijc.30369.
328. Green, R. *et al.* Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. *Fam. Cancer* **6**, 53–62 (2007) doi: 10.1007/s10689-006-9104-x.
329. Woods, M. O. *et al.* The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut* **59**, 1369–1377 (2010) doi:10.1136/gut.2010.208462.
330. Wang, P. P. *et al.* Validity of random-digit-dialing in recruiting controls in a case-control study. *Am J Health Behav* **33**, 513–520 (2009) doi: 10.5993/ajhb.33.5.4.
331. Schemper, M. & Smith, T. L. A note on quantifying follow-up in studies of failure time. *Control. Clin. Trials* **17**, 343–346 (1996) doi: 10.1016/0197-2456(96)00075-x.
332. Rodriguez, S., Gaunt, T. R. & Day, I. N. M. Hardy-Weinberg equilibrium testing of biological ascertainment for Mendelian randomization studies. *Am. J. Epidemiol.* **169**, 505–514 (2009) doi:10.1093/aje/kwn359.
333. Warnes G, with contributions from Gorjanc G, Leisch F and Man M. *genetics: Population Genetics. R package version 1.3.8.1*. (2013) Available at: https://cran.r-project.org/web/packages/genetics/index.html (accessed on Nov 23, 2021).

334. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723 (1974) doi: 10.1109/TAC.1974.1100705.

335. Jones, S. *et al.* Somatic mutations in the chromatin remodeling gene ARID1A occur in several tumor types. *Hum. Mutat.* **33**, 100–103 (2012) doi:10.1002/humu.21633.

336. Li, F. & Lai, M. Colorectal cancer, one entity or three. *J. Zhejiang Univ. Sci. B* **10**, 219–229 (2009) doi:10.1631/jzus.B0820273.

337. Marrett, L. D., De, P., Airia, P., Dryer, D. & Steering Committee of Canadian Cancer Statistics 2008. Cancer in Canada in 2008. *CMAJ Can. Med. Assoc. J. J. Assoc. Medicale Can.* **179**, 1163–1170 (2008) doi:10.1503/cmaj.080760.

338. Savas, S. & Liu, G. Genetic variations as cancer prognostic markers: review and update. *Hum. Mutat.* **30**, 1369–1377 (2009) doi:10.1002/humu.21078.

339. Verma, M. Personalized medicine and cancer. *J. Pers. Med.* **2**, 1–14 (2012) doi:10.3390/jpm2010001.

340. Mills, R. E. *et al.* An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res.* **16**, 1182–1190 (2006) doi: 10.1101/gr.4565806.

341. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006) doi:10.1038/nature05329.

342. Lupski, J. R. Genomic rearrangements and sporadic disease. *Nat. Genet.* **39**, S43-47 (2007) doi:10.1038/ng2084.

343. Low, J. S. Y. *et al.* A Genome Wide Study of Copy Number Variation Associated with Nasopharyngeal Carcinoma in Malaysian Chinese Identifies CNVs at 11q14.3 and 6p21.3 as Candidate Loci. *PloS One* **11**, e0145774 (2016) doi:10.1371/journal.pone.0145774.

344. Tervasmäki, A., Winqvist, R., Jukkola-Vuorinen, A. & Pylkäs, K. Recurrent CYP2C19 deletion allele is associated with triple-negative breast cancer. *BMC Cancer* **14**, 902 (2014) doi:10.1186/1471-2407-14-902.

345. Torres, F., Barbosa, M. & Maciel, P. Recurrent copy number variations as risk factors for neurodevelopmental disorders: critical overview and analysis of clinical implications. *J. Med. Genet.* **53**, 73–90 (2016) doi:10.1136/jmedgenet-2015-103366.

346. Li, Z. *et al.* A genome-wide assessment of rare copy number variants in colorectal cancer. *Oncotarget* **6**, 26411–26423 (2015) doi:10.18632/oncotarget.4621.

347. Masson, A. L. *et al.* Copy number variation in hereditary non-polyposis colorectal cancer. *Genes* **4**, 536–555 (2013) doi:10.3390/genes4040536.

348. Talseth-Palmer, B. A. *et al.* Continuing difficulties in interpreting CNV data: lessons from a genome-wide CNV association study of Australian HNPCC/lynch syndrome patients. *BMC Med. Genomics* **6**, 10 (2013) doi:10.1186/1755-8794-6-10.

349. Venkatachalam, R. *et al.* Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int. J. Cancer* **129**, 1635–1642 (2011) doi:10.1002/ijc.25821.

350. Holley, S. L. *et al.* Polymorphisms in the glutathione S-transferase mu cluster are associated with tumour progression and patient outcome in colorectal cancer. *Int. J. Oncol.* **28**, 231–236 (2006) doi: 10.3892/ijo.28.1.231.

351. Kap, E. J. *et al.* Genetic variants in the glutathione S-transferase genes and survival in colorectal cancer patients after chemotherapy and differences according to treatment with oxaliplatin. *Pharmacogenet. Genomics* **24**, 340–347 (2014) doi:10.1097/FPC.0000000000000059.

352. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007) doi: 10.1093/nar/gkm076.

353. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007) doi: 10.1101/gr.6861907.

354. Hubbard, T. J. P. *et al.* Ensembl 2009. *Nucleic Acids Res.* **37**, D690-697 (2009) doi:10.1093/nar/gkn828.

355. Mi, H. & Thomas, P. PANTHER pathway: an ontology-based pathway database coupled with data analysis tools. *Methods Mol. Biol. Clifton NJ* **563**, 123–140 (2009) doi:10.1007/978-1-60761-175-2_7.

356. Therneau, T. M. & Grambsch, P. M. *Modeling Survival Data: Extending the Cox Model*. (Springer New York, 2000). doi:10.1007/978-1-4757-3294-8.

357. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010) doi: 10.1038/nature09534.

358. Campbell, C. D. *et al.* Population-genetic properties of differentiated human copy-number polymorphisms. *Am. J. Hum. Genet.* **88**, 317–332 (2011) doi:10.1016/j.ajhg.2011.02.004.

359. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010) doi:10.1038/nature08516.

360. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.* **29**, 512–520 (2011) doi:10.1038/nbt.1852.

361. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Brief. Funct. Genomic. Proteomic.* **8**, 353–366 (2009) doi:10.1093/bfgp/elp017.

362. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010) doi:10.1038/nature09146.

363. Itsara, A. *et al.* Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Hum. Genet.* **84**, 148–161 (2009) doi:10.1016/j.ajhg.2008.12.014.

364. Bhat, T. A. & Singh, R. P. Tumor angiogenesis--a potential target in cancer chemoprevention. *Food Chem. Toxicol. Int. J. Publ. Br. Ind. Biol. Res. Assoc.* **46**, 1334–1345 (2008) doi:10.1016/j.fct.2007.08.032.

365. Liotta, L. A., Steeg, P. S. & Stetler-Stevenson, W. G. Cancer metastasis and angiogenesis: an imbalance of positive and negative regulation. *Cell* **64**, 327–336 (1991) doi:10.1016/0092-8674(91)90642-c.

366. Liu, X., Ji, Q., Fan, Z. & Li, Q. Cellular signaling pathways implicated in metastasis of colorectal cancer and the associated targeted agents. *Future Oncol. Lond. Engl.* **11**, 2911–2922 (2015) doi:10.2217/fon.15.235.

367. Zhan, T., Rindtorff, N. & Boutros, M. Wnt signaling in cancer. *Oncogene* **36**, 1461–1473 (2017) doi:10.1038/onc.2016.304.

368. Dominguez-Valentin, M., Therkildsen, C., Da Silva, S. & Nilbert, M. Familial colorectal cancer type X: genetic profiles and phenotypic features. *Mod. Pathol. Off. J. U. S. Can. Acad. Pathol. Inc* **28**, 30–36 (2015) doi:10.1038/modpathol.2014.49.

369. Nieminen, T. T. *et al.* Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch

repair deficiency. *Gastroenterology* **147**, 595-598.e5 (2014) doi:10.1053/j.gastro.2014.06.009.

370. Sánchez-Tomé, E. *et al.* Genome-wide linkage analysis and tumoral characterization reveal heterogeneity in familial colorectal cancer type X. *J. Gastroenterol.* **50**, 657–666 (2015) doi:10.1007/s00535-014-1009-0.

371. Schulz, E. *et al.* Germline variants in the SEMA4A gene predispose to familial colorectal cancer type X. *Nat. Commun.* **5**, 5191 (2014) doi:10.1038/ncomms6191.

372. Griffin, M. R., Bergstralh, E. J., Coffey, R. J., Beart, R. W. & Melton, L. J. Predictors of survival after curative resection of carcinoma of the colon and rectum. *Cancer* **60**, 2318–2324 (1987) doi:10.1002/1097-0142(19871101)60:9<2318::aid-cncr2820600934>3.0.co;2-b.

373. Derwinger, K., Kodeda, K. & Gerjy, R. Age aspects of demography, pathology and survival assessment in colorectal cancer. *Anticancer Res.* **30**, 5227–5231 (2010) PMID: 21187518.

374. Ganapathi, S. *et al.* Colorectal cancer in the young: trends, characteristics and outcome. *Int. J. Colorectal Dis.* **26**, 927 (2011) doi: 10.1007/s00384-011-1174-z.

375. O'Connell, J. B., Maggard, M. A., Liu, J. H., Etzioni, D. A. & Ko, C. Y. Are survival rates different for young and older patients with rectal cancer? *Dis. Colon Rectum* **47**, 2064–2069 (2004) doi: 10.1007/s10350-004-0738-1.

376. Aghili, M., Izadi, S., Madani, H. & Mortazavi, H. Clinical and pathological evaluation of patients with early and late recurrence of colorectal cancer. *Asia Pac. J. Clin. Oncol.* **6**, 35–41 (2010) doi:10.1111/j.1743-7563.2010.01275.x.

377. Gatza, C. E., Oh, S. Y. & Blobe, G. C. Roles for the type III TGF-beta receptor in human cancer. *Cell. Signal.* **22**, 1163–1174 (2010) doi:10.1016/j.cellsig.2010.01.016.

378. Gatza, C. E. *et al.* Type III TGF-β receptor enhances colon cancer cell migration and anchorage-independent growth. *Neoplasia N. Y. N* **13**, 758–770 (2011) doi:10.1593/neo.11528.

379. Kwon, M. *et al.* Filamin A interacting protein 1-like inhibits WNT signaling and MMP expression to suppress cancer cell invasion and metastasis. *Int. J. Cancer* **135**, 48–60 (2014) doi:10.1002/ijc.28662.

380. Kwon, M. *et al.* Down-regulation of Filamin A interacting protein 1-like Is associated with promoter methylation and an invasive phenotype in breast, colon, lung and pancreatic cancers [corrected]. *PloS One* **8**, e82620 (2013) doi:10.1371/journal.pone.0082620.

381. Park, Y. L. *et al.* Filamin A interacting protein 1-like expression inhibits progression in colorectal cancer. *Oncotarget* **7**, 72229–72241 (2016) doi:10.18632/oncotarget.12664.

382. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012) doi:10.1038/nature11247.

383. Cortese, G., Scheike, T. H. & Martinussen, T. Flexible survival regression modelling. *Stat. Methods Med. Res.* **19**, 5–28 (2010) doi: 10.1177/0962280209105022.

384. Sargent, D. *et al.* Evidence for cure by adjuvant therapy in colon cancer: observations based on individual patient data from 20,898 patients on 18 randomized trials. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **27**, 872–877 (2009) doi:10.1200/JCO.2008.19.5362.

385. Ryuk, J. P. *et al.* Predictive factors and the prognosis of recurrence of colorectal cancer within 2 years after curative resection. *Ann. Surg. Treat. Res.* **86**, 143–151 (2014) doi: 10.4174/astr.2014.86.3.143.

386. Marzouk, O. & Schofield, J. Review of histopathological and molecular prognostic features in colorectal cancer. *Cancers* **3**, 2767–2810 (2011) doi: 10.3390/cancers3022767.

387. Popat, S., Hubner, R. & Houlston, R. S. Systematic review of microsatellite instability and colorectal cancer prognosis. *J. Clin. Oncol.* **23**, 609–618 (2005) doi: 10.1200/JCO.2005.01.086.

388. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954 (2002) doi:10.1038/nature00766.

389. Phipps, A. I. *et al.* Family history of colorectal cancer is not associated with colorectal cancer survival regardless of microsatellite instability status. *Cancer Epidemiol. Biomark. Prev. Publ. Am. Assoc. Cancer Res. Cosponsored Am. Soc. Prev. Oncol.* **23**, 1700–1704 (2014) doi:10.1158/1055-9965.EPI-14-0533.

390. Johns, L. E. & Houlston, R. S. A systematic review and meta-analysis of familial colorectal cancer risk. *Am. J. Gastroenterol.* **96**, 2992–3003 (2001) doi: 10.1111/j.1572-0241.2001.04677.x.

391. Karran, P. Microsatellite instability and DNA mismatch repair in human cancer. *Semin. Cancer Biol.* **7**, 15–24 (1996) doi: 10.1006/scbi.1996.0003.

392. Ardekani, G. S., Jafarnejad, S. M., Tan, L., Saeedi, A. & Li, G. The prognostic value of BRAF mutation in colorectal cancer and melanoma: a systematic review and meta-analysis. *PloS One* **7**, e47054 (2012) doi: 10.1371/journal.pone.0047054.

393. Natarajan, L. *et al.* Time-varying effects of prognostic factors associated with disease-free survival in breast cancer. *Am. J. Epidemiol.* **169**, 1463–1470 (2009) doi: 10.1093/aje/kwp077.

394. Sigounas, D. E., Tatsioni, A., Christodoulou, D. K., Tsianos, E. V. & Ioannidis, J. P. New prognostic markers for outcome of acute pancreatitis: overview of reporting in 184 studies. *Pancreas* **40**, 522–532 (2011) doi: 10.1097/MPA.0b013e31820bf8ac.

395. Werdyani, S. *et al.* Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. *Cancer Med.* **6**, 1220–1232 (2017) doi: 10.1002/cam4.1074.

396. Yu, Y., Cheng, D., Parfrey, P., Liu, G. & Savas, S. Two functional indel polymorphisms in the promoter region of the Brahma gene (BRM) and disease risk and progression-free survival in colorectal cancer. *PLoS One* **13**, e0198873 (2018) doi: 10.1371/journal.pone.0198873.

397. Farewell, V. T. The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041–1046 (1982) doi: 10.2307/2529885.

398. Berian, J. R., Benson III, A. B. & Nelson, H. Young age and aggressive treatment in Colon Cancer. *Jama* **314**, 613–614 (2015) doi: 10.1001/jama.2015.9379.

399. Law, J.-H., Koh, F. H. & Tan, K.-K. Young colorectal cancer patients often present too late. *Int. J. Colorectal Dis.* **32**, 1165–1169 (2017) doi: 10.1007/s00384-017-2837-1.

400. McArdle, C., McMillan, D. & Hole, D. Male gender adversely affects survival following surgery for colorectal cancer. *Br. J. Surg.* **90**, 711–715 (2003) doi: 10.1002/bjs.4098.

401. Wichmann, M. *et al.* Gender differences in long-term survival of patients with colorectal cancer. *Br. J. Surg.* **88**, 1092–1098 (2001) doi: 10.1046/j.0007-1323.2001.01819.x.

402. Yang, Y. *et al.* Gender differences in colorectal cancer survival: A meta-analysis. *Int. J. Cancer* **141**, 1942–1949 (2017) doi: 10.1002/ijc.30827.

403. Wang, W.-S. *et al.* Preoperative carcinoembryonic antigen level as an independent prognostic factor in colorectal cancer: Taiwan experience. *Jpn. J. Clin. Oncol.* **30**, 12–16 (2000) doi: 10.1093/jjco/hyd003.

404. Bertario, L. *et al.* Survival of patients with hereditary colorectal cancer: comparison of HNPCC and colorectal cancer in FAP patients with sporadic colorectal cancer. *Int. J. Cancer* **80**, 183–187 (1999) doi: 10.1002/(sici)1097-0215(19990118)80:2<183::aid-ijc4>3.0.co;2-w.

405. Butterworth, A. S., Higgins, J. P. & Pharoah, P. Relative and absolute risk of colorectal cancer for individuals with a family history: a meta-analysis. *Eur. J. Cancer* **42**, 216–227 (2006) doi: 10.1016/j.ejca.2005.09.023.

406. Malesci, A. *et al.* Reduced likelihood of metastases in patients with microsatellite-unstable colorectal cancer. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **13**, 3831–3839 (2007) doi: 10.1158/1078-0432.CCR-07-0366.

407. Mayo, S. C. *et al.* Refining the definition of perioperative mortality following hepatectomy using death within 90 days as the standard criterion. *HPB* **13**, 473–482 (2011) doi: 10.1111/j.1477-2574.2011.00326.x.

408. Rutegård, M., Haapamäki, M., Matthiessen, P. & Rutegård, J. Early postoperative mortality after surgery for rectal cancer in Sweden, 2000–2011. *Colorectal Dis.* **16**, 426–432 (2014) doi: 10.1111/codi.12572.

409. van Eeghen, E. E., den Boer, F. C. & Loffeld, R. J. Thirty days post-operative mortality after surgery for colorectal cancer: a descriptive study. *J. Gastrointest. Oncol.* **6**, 613–617 (2015) doi:10.3978/j.issn.2078-6891.2015.079.

410. Obrand, D. I. & Gordon, P. H. Incidence and patterns of recurrence following curative resection for colorectal carcinoma. *Dis. Colon Rectum* **40**, 15–24 (1997) doi: 10.1007/BF02055676.

411. Pugh, S. A. *et al.* Site and stage of colorectal cancer influence the likelihood and distribution of disease recurrence and postrecurrence survival: data from the FACS randomized controlled trial. *Ann. Surg.* **263**, 1143–1147 (2016) doi: 10.1097/SLA.0000000000001351.

412. Singer, G. *et al.* Mutations in BRAF and KRAS characterize the development of low-grade ovarian serous carcinoma. *J. Natl. Cancer Inst.* **95**, 484–486 (2003) doi: 10.1093/jnci/95.6.484.

413. Xing, M. *et al.* Association between BRAF V600E mutation and recurrence of papillary thyroid cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **33**, 42–50 (2015) doi:10.1200/JCO.2014.56.8253.

414. Sánchez-Torres, J. M., Viteri, S., Molina, M. A. & Rosell, R. BRAF mutant non-small cell lung cancer and treatment with BRAF inhibitors. *Transl. Lung Cancer Res.* **2**, 244–250 (2013) doi: 10.3978/j.issn.2218-6751.2013.04.01.

415. Huang, D. *et al.* Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer Metastasis Rev.* **37**, 173–187 (2018) doi: 10.1007/s10555-017-9726-5.

416. Palomba, G. *et al.* Prognostic impact of KRAS, NRAS, BRAF, and PIK3CA mutations in primary colorectal carcinomas: a population-based study. *J. Transl. Med.* **14**, 292 (2016) doi: 10.1186/s12967-016-1053-z.

417. Won, D. D. *et al.* The prognostic significance of KRAS and BRAF mutation status in Korean colorectal cancer patients. *BMC Cancer* **17**, 403 (2017) doi: 10.1186/s12885-017-3381-7.

418. Neugut, A. I. *et al.* Duration of adjuvant chemotherapy for colon cancer and survival among the elderly. *J Clin Oncol* **24**, 2368–2375 (2006) doi: 10.1200/JCO.2005.04.5005.

419. Demicheli, R. *et al.* Breast cancer recurrence dynamics following adjuvant CMF is consistent with tumor dormancy and mastectomy-driven acceleration of the metastatic process. *Ann. Oncol.* **16**, 1449–1457 (2005) doi: 10.1093/annonc/mdi280.

420. Dignam, J. J. *et al.* Hazard of recurrence and adjuvant treatment effects over time in lymph node-negative breast cancer. *Breast Cancer Res. Treat.* **116**, 595–602 (2009) doi: 10.1007/s10549-008-0200-5.

421. Jatoi, I. *et al.* Time-varying effects of breast cancer adjuvant systemic therapy. *JNCI J. Natl. Cancer Inst.* **108**, djv304 (2016) doi: 10.1093/jnci/djv304.

422. Sofia Vala, I. *et al.* Low doses of ionizing radiation promote tumor growth and metastasis by enhancing angiogenesis. *PloS One* **5**, e11222 (2010) doi:10.1371/journal.pone.0011222.

423. Sundahl, N., Duprez, F., Ost, P., De Neve, W. & Mareel, M. Effects of radiation on the metastatic process. *Mol. Med. Camb. Mass* **24**, 16 (2018) doi:10.1186/s10020-018-0015-8.

424. Vilalta, M., Rafat, M. & Graves, E. E. Effects of radiation on metastasis and tumor cell migration. *Cell. Mol. Life Sci. CMLS* **73**, 2999–3007 (2016) doi:10.1007/s00018-016-2210-5.

425. Song, J. W. & Chung, K. C. Observational studies: cohort and case- control studies. *Plast. Reconstr. Surg.* **126**, 2234–2242 (2010) doi:10.1097/PRS.0b013e3181f44abc.

426. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA. Cancer J. Clin.* **68**, 394–424 (2018) doi:10.3322/caac.21492.

427. Li, M., Li, J. Y., Zhao, A. L. & Gu, J. Colorectal cancer or colon and rectal cancer? Clinicopathological comparison between colonic and rectal carcinomas. *Oncology* **73**, 52–57 (2007) doi: 10.1159/000120628.

428. Wolpin, B. M., Meyerhardt, J. A., Mamon, H. J. & Mayer, R. J. Adjuvant treatment of colorectal cancer. *CA. Cancer J. Clin.* **57**, 168–185 (2007) doi: 10.3322/canjclin.57.3.168.

429. He, Y. *et al.* A comprehensive study of the effect on colorectal cancer survival of common germline genetic variation previously linked with cancer prognosis. *Cancer Epidemiol. Biomarkers Prev.* **28**, 1944–1946 (2019) doi:10.1158/1055-9965.EPI-19-0596.

430. Riera, P. *et al.* Genetic variants in the VEGF pathway as prognostic factors in stages II and III colon cancer. *Pharmacogenomics J.* **18**, 556–564 (2018) doi:10.1038/s41397-017-0009-x.

431. Savas, S. *et al.* A survival association study of 102 polymorphisms previously associated with survival outcomes in colorectal cancer. *BioMed Res. Int.* **2015**, 9 (2015) doi: 10.1155/2015/968743.

432. Sebio, A. *et al.* Genetic variants within obesity-related genes are associated with tumor recurrence in patients with stages II/III colon cancer. *Pharmacogenet. Genomics* **25**, 30–37 (2015) doi:10.1097/FPC.0000000000000101.

433. Theodoratou, E. *et al.* Genome-wide scan of the effect of common nsSNPs on colorectal cancer survival outcome. *Br. J. Cancer* **119**, 988–993 (2018) doi:10.1038/s41416-018-0117-7.

434. Penney, K. L. *et al.* Genetic Variant Associated with Survival of Patients with Stage II-III Colon Cancer. *Clin. Gastroenterol. Hepatol.* (2019) doi:10.1016/j.cgh.2019.11.046.

435. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012) doi:10.1038/nmeth.1785.

436. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009) doi:10.1371/journal.pgen.1000529.

437. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007) doi: 10.1086/519795.

438. Owzar, K., Li, Z., Cox, N. & Jung, S.-H. Power and sample size calculations for SNP association studies with censored time-to-event outcomes. *Genet. Epidemiol.* **36**, 538–548 (2012) doi:10.1002/gepi.21645.

439. Turner, S. D. qqman: an R package for visualizing GWAS results using QQ and manhattan plots. *J. Open Source Softw.* **3**, 731 (2018) doi: 10.21105/joss.00731.

440. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010) doi: 10.1093/bioinformatics/btq419.

441. Grossman, R. L. *et al.* Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016) doi:10.1056/NEJMp1607591.

442. Liu, J. *et al.* An integrated TCGA Pan-Cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018) doi:10.1016/j.cell.2018.02.052.

443. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930-934 (2012) doi:10.1093/nar/gkr917.

444. Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020) doi:10.1038/s41587-020-0546-8.

445. Guinney, J. *et al.* The consensus molecular subtypes of colorectal cancer. *Nat. Med.* **21**, 1350–1356 (2015) doi:10.1038/nm.3967.

446. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012) doi:10.1101/gr.137323.112.

447. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013) doi:10.1038/ng.2653.

448. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013) doi:10.1038/ng.2764.

449. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39**, D52-57 (2011) doi:10.1093/nar/gkq1237.

450. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002) doi:10.1101/gr.229102.

451. Savas, S. & Younghusband, H. B. dbCPCO: a database of genetic markers tested for their predictive and prognostic value in colorectal cancer. *Hum. Mutat.* **31**, 901–907 (2010) doi:10.1002/humu.21285.

452. Llorian, M., Beullens, M., Andrés, I., Ortiz, J.-M. & Bollen, M. SIPP1, a novel pre-mRNA splicing factor and interactor of protein phosphatase-1. *Biochem. J.* **378**, 229–238 (2004) doi:10.1042/BJ20030950.

453. Wang, L. *et al.* The miR-29c-KIAA1199 axis regulates gastric cancer migration by binding with WBP11 and PTP4A3. *Oncogene* **38**, 3134–3150 (2019) doi: 10.1038/s41388-018-0642-0.

454. Segditsas, S. & Tomlinson, I. Colorectal cancer and genetic alterations in the Wnt pathway. *Oncogene* **25**, 7531–7537 (2006) doi: 10.1038/sj.onc.1210059.

455. Zhang, Z. *et al.* BRM/SMARCA2 promotes the proliferation and chemoresistance of pancreatic cancer cells by targeting JAK2/STAT3 signaling. *Cancer Lett.* **402**, 213–224 (2017) doi:10.1016/j.canlet.2017.05.006.

456. Zhu, Y. *et al.* Brahma regulates the Hippo pathway activity through forming complex with Yki-Sd and regulating the transcription of Crumbs. *Cell. Signal.* **27**, 606–613 (2015) doi:10.1016/j.cellsig.2014.12.002.

457. Fang, R. *et al.* Inactivation of BRM/SMARCA2 sensitizes clear cell renal cell carcinoma to histone deacetylase complex inhibitors. *Pathol. Res. Pract.* **216**, 152867 (2020) doi:10.1016/j.prp.2020.152867.

458. Jancewicz, I., Siedlecki, J. A., Sarnowski, T. J. & Sarnowska, E. BRM: the core ATPase subunit of SWI/SNF chromatin-remodelling complex-a tumour suppressor or tumour-promoting factor? *Epigenetics Chromatin* **12**, 68 (2019) doi:10.1186/s13072-019-0315-4.

459. Marquez-Vilendrer, S. B., Rai, S. K., Gramling, S. J., Lu, L. & Reisman, D. N. Loss of the SWI/SNF ATPase subunits BRM and BRG1 drives lung cancer development. *Oncoscience* **3**, 322–336 (2016) doi:10.18632/oncoscience.323.

460. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001-1006 (2014) doi:10.1093/nar/gkt1229.

461. Bigagli, E. *et al.* DNA copy number alterations, gene expression changes and disease-free survival in patients with colorectal cancer: a 10 year follow-up. *Cell. Oncol.* **39**, 545–558 (2016) doi:10.1007/s13402-016-0299-z.

462. Wang, H., Liang, L., Fang, J.-Y. & Xu, J. Somatic gene copy number alterations in colorectal cancer: new quest for cancer drivers and biomarkers. *Oncogene* **35**, 2011–2019 (2016) doi:10.1038/onc.2015.304.

463. Zaidi, S. H. *et al.* Landscape of somatic single nucleotide variants and indels in colorectal cancer and impact on survival. *Nat. Commun.* **11**, 3644 (2020) doi:10.1038/s41467-020-17386-z.

464. Ionita-Laza, I., Rogers, A. J., Lange, C., Raby, B. A. & Lee, C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics* **93**, 22–26 (2009) doi:10.1016/j.ygeno.2008.08.012.

465. Gamazon, E. R. & Stranger, B. E. The impact of human copy number variation on gene expression. *Brief. Funct. Genomics* **14**, 352–357 (2015) doi:10.1093/bfgp/elv017.

466. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009) doi:10.1038/nature08494.

467. Abadi, A., Saadat, S., Yavari, P., Bajdik, C. & Jalili, P. Comparison of Aalen's additive and Cox proportional hazards models for breast cancer survival: analysis of population- based data from British Columbia, Canada. *Asian Pac. J. Cancer Prev. APJCP* **12**, 3113–3116 (2011) PMID: 22393999.

468. Ataee Dizaji, P., Vasheghani Farahani, M., Sheikhaliyan, A. & Biglarian, A. Application of additive hazards models for analyzing survival of breast cancer patients. *J. Res. Med. Sci.* **25**, 99 (2020) doi:10.4103/jrms.JRMS_701_19.

469. Baulies, S. *et al.* Time-varying effect and long-term survival analysis in breast cancer patients treated with neoadjuvant chemotherapy. *Br. J. Cancer* **113**, 30 (2015) doi: 10.1038/bjc.2015.174.

470. Chen, Q. *et al.* Time-varying effects of FOXA1 on breast cancer prognosis. *Breast Cancer Res. Treat.* (2021) doi:10.1007/s10549-021-06125-7.

471. Hilsenbeck, S. G. *et al.* Time-dependence of hazard ratios for prognostic factors in primary breast cancer. *Breast Cancer Res. Treat.* **52**, 227–237 (1998) doi:10.1023/A:1006133418245.

472. Mazroui, Y. *et al.* Time-varying coefficients in a multivariate frailty model: Application to breast cancer recurrences of several types and death. *Lifetime Data Anal.* **22**, 191–215 (2016) doi:10.1007/s10985-015-9327-y.

473. Perperoglou, A., Keramopoullos, A. & van Houwelingen, H. C. Approaches in modelling long-term survival: an application to breast cancer. *Stat. Med.* **26**, 2666–2685 (2007) doi:10.1002/sim.2729.

474. Rakovitch, E. *et al.* The time-varying effect of radiotherapy after breast-conserving surgery for DCIS. *Breast Cancer Res. Treat.* **178**, 221–230 (2019) doi:10.1007/s10549-019-05377-8.

475. Rogoz, B., Houzé de l'Aulnoit, A., Duhamel, A. & Houzé de l'Aulnoit, D. Thirty-Year Trends of Survival and Time-Varying Effects of Prognostic Factors in Patients With Metastatic Breast Cancer—A Single Institution Experience. *Clin. Breast Cancer* **18**, 246–253 (2018) doi:10.1016/j.clbc.2017.08.012.

476. Schmitt, M. *et al.* Time-varying prognostic impact of tumour biological factors urokinase (uPA), PAI-1 and steroid hormone receptor status in primary breast cancer. *Br. J. Cancer* **76**, 306–311 (1997) doi:10.1038/bjc.1997.383.

477. Solomayer, E. F. *et al.* Time independence of the prognostic impact of tumor cell detection in the bone marrow of primary breast cancer patients. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **7**, 4102–4108 (2001) PMID: 11751508.

478. Warwick, J., Tabàr, L., Vitak, B. & Duffy, S. W. Time-dependent effects on survival in breast carcinoma: Results of 20 years of follow-up from the Swedish two-county study. *Cancer* **100**, 1331–1336 (2004) doi:10.1002/cncr.20140.

479. Zhang, M. *et al.* Time-varying effects of prognostic factors associated with long-term survival in breast cancer. *Endocr. Relat. Cancer* **25**, 509–521 (2018) doi:10.1530/ERC-17-0502.

480. Shida, D. *et al.* Prognostic impact of primary tumor location in Stage III colorectal cancer-right-sided colon versus left-sided colon versus rectum: a nationwide multicenter retrospective study. *J. Gastroenterol.* **55**, 958–968 (2020) doi:10.1007/s00535-020-01706-7.

481. Lee, Y.-C., Lee, Y.-L., Chuang, J.-P. & Lee, J.-C. Differences in survival between colon and rectal cancer from SEER data. *PloS One* **8**, e78709 (2013) doi:10.1371/journal.pone.0078709.

482. Moghimi-Dehkordi, B. *et al.* Statistical comparison of survival models for analysis of cancer data. *Asian Pac. J. Cancer Prev. APJCP* **9**, 417–420 (2008) PMID: 18990013.

483. Pourhoseingholi, M. A. *et al.* Comparing Cox regression and parametric models for survival of patients with gastric carcinoma. *Asian Pac. J. Cancer Prev. APJCP* **8**, 412–416 (2007) PMID: 18159979.

APPENDICES

Appendix A: The latest ethics approval of my research and copyright permissions for the use of figures and tables from published papers (Chapter 1)

Latest ethics approval of my research:

MEMORIAL UNIVERSITY

Yu, Yajun <yy6084@mun.ca>

HREB - Approval of Ethics Renewal 553028

1 message

administrator@hrea.ca <administrator@hrea.ca> Thu, Nov 4, 2021 at 9:24 AM
To: "Yu Yajun(Principal Investigator)" <yy6084@mun.ca>
Cc: "Savas Sevtap(Supervisor)" <savas@mun.ca>, administrator@hrea.ca

Researcher Portal File #: 20161668

Dear Mr. Yajun Yu:

This e-mail serves as notification that your ethics renewal for study HREB # 2015.294 – Genetic and epidemiological studies investigating the susceptibility and outcome in colorectal cancer – has been **approved**. Please log in to the Researcher Portal to view the approved event.

Ethics approval for this project has been granted for a period of twelve months effective from December 1, 2021 to December 1, 2022.

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

The ethics renewal [will be reported] to the Health Research Ethics Board at their meeting dated November 18, 2021.

Thank you,

Research Ethics Office

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

Copyright permission for Fig. 1.1:

ELSEVIER LICENSE
TERMS AND CONDITIONS

Sep 16, 2021

This Agreement between Mr. Yajun Yu ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5151090213330
License date	Sep 16, 2021
Licensed Content Publisher	Elsevier
Licensed Content Publication	Critical Reviews in Oncology/Hematology
Licensed Content Title	Role of hypoxia-inducible factors (HIF) in the maintenance of stemness and malignancy of colorectal cancer
Licensed Content Author	Ramakrishna Vadde,Sarojamma Vemula,Rajeswari Jinka,Neha Merchant,Pallaval Veera Bramhachari,Ganji Purnachandra Nagaraju
Licensed Content Date	May 1, 2017
Licensed Content Volume	113
Licensed Content Issue	n/a

Licensed Content Pages	6
Start Page	22
End Page	27
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	Genetic and clinico-demographic factors with and without time-varying associations with survival outcomes in colorectal cancer
Institution name	Memorial University of Newfoundland
Expected presentation date	Nov 2021
Portions	Fig. 1
	Mr. Yajun Yu Craig L. Dobbin Genetics Research Centre

Requestor Location
St. John's, NL A1B 3V6
Canada
Attn: Mr. Yajun Yu

Publisher Tax ID GB 494 6272 12

Total 0.00 CAD

Terms and Conditions

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier

Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of

articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.

Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's posting policy for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.10

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Copyright permission for Fig. 1.2:

ELSEVIER LICENSE
TERMS AND CONDITIONS

Sep 23, 2021

This Agreement between Mr. Yajun Yu ("You") and Elsevier ("Elsevier") consists of your license details and the terms and conditions provided by Elsevier and Copyright Clearance Center.

License Number	5154880332833
License date	Sep 23, 2021
Licensed Content Publisher	Elsevier
Licensed Content Publication	Gastroenterology
Licensed Content Title	Genetics and Genetic Biomarkers in Sporadic Colorectal Cancer
Licensed Content Author	John M. Carethers,Barbara H. Jung
Licensed Content Date	Oct 1, 2015
Licensed Content Volume	149
Licensed Content Issue	5
Licensed Content Pages	17
Start Page	1177
End Page	1190.e3

Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1
Format	both print and electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Title	Genetic and clinico-demographic factors with and without time-varying associations with survival outcomes in colorectal cancer
Institution name	Memorial University of Newfoundland
Expected presentation date	Nov 2021
Portions	Figure 1. (note that only the subfigures B and C of Figure 1 will be used)
Requestor Location	Mr. Yajun Yu Craig L. Dobbin Genetics Research Centre St. John's, NL A1B 3V6 Canada Attn: Mr. Yajun Yu
Publisher Tax ID	GB 494 6272 12
Total	0.00 CAD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at http://myaccount.copyright.com).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier's permissions helpdesk here). No modifications can be made to any Lancet figures/tables and they must be reproduced in full.

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. **Translation**: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article.

16. **Posting licensed content on any Website**: The following terms and conditions apply as follows: Licensing material from an Elsevier journal: All content posted to the web site must maintain the copyright information line on the bottom of each image; A hyper-text must be included to the Homepage of the journal from which you are licensing at http://www.sciencedirect.com/science/journal/xxxxx or the Elsevier homepage for books at http://www.elsevier.com; Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

Licensing material from an Elsevier book: A hyper-text link must be included to the Elsevier homepage at http://www.elsevier.com . All content posted to the web site must maintain the copyright information line on the bottom of each image.

Posting licensed content on Electronic reserve: In addition to the above the following clauses are applicable: The web site must be password-protected and made available only to bona fide students registered on a relevant course. This permission is granted for 1 year only. You may obtain a new license for future website posting.

17. **For journal authors:** the following clauses are applicable in addition to the above:

Preprints:

A preprint is an author's own write-up of research results and analysis, it has not been peer-reviewed, nor has it had any other value added to it by a publisher (such as formatting, copyright, technical enhancement etc.).

Authors can share their preprints anywhere at any time. Preprints should not be added to or enhanced in any way in order to appear more like, or to substitute for, the final versions of articles however authors can update their preprints on arXiv or RePEc with their Accepted Author Manuscript (see below).

If accepted for publication, we encourage authors to link from the preprint to their formal publication via its DOI. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help users to find, access, cite and use the best available version. Please note that Cell Press, The Lancet and some society-owned have different preprint policies. Information on these policies is available on the journal homepage.

Accepted Author Manuscripts: An accepted author manuscript is the manuscript of an article that has been accepted for publication and which typically includes author-incorporated changes suggested during submission, peer review and editor-author communications.

Authors can share their accepted author manuscript:

- immediately
  - via their non-commercial person homepage or blog
  - by updating a preprint in arXiv or RePEc with the accepted manuscript
  - via their research institute or institutional repository for internal institutional uses or as part of an invitation-only research collaboration work-group
  - directly by providing copies to their students or to research collaborators for their personal use
  - for private scholarly sharing as part of an invitation-only work group on commercial sites with which Elsevier has an agreement
- After the embargo period
  - via non-commercial hosting platforms such as their institutional repository
  - via commercial sites with which Elsevier has an agreement

In all cases accepted manuscripts should:

- link to the formal publication via its DOI
- bear a CC-BY-NC-ND license - this is easy to do
- if aggregated with other manuscripts, for example in a repository or other site, be shared in alignment with our hosting policy not be added to or enhanced in any way to appear more like, or to substitute for, the published journal article.

Published journal article (JPA): A published journal article (PJA) is the definitive final record of published research that appears or will appear in the journal and embodies all

value-adding publishing activities including peer review co-ordination, copy-editing, formatting, (if relevant) pagination and online enrichment.

Policies for sharing publishing journal articles differ for subscription and gold open access articles:

Subscription Articles: If you are an author, please share a link to your article rather than the full-text. Millions of researchers have access to the formal publications on ScienceDirect, and so links will help your users to find, access, cite, and use the best available version.

Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

If you are affiliated with a library that subscribes to ScienceDirect you have additional private sharing rights for others' research accessed under that agreement. This includes use for classroom teaching and internal training at the institution (including use in course packs and courseware programs), and inclusion of the article for grant funding purposes.

Gold Open Access Articles: May be shared according to the author-selected end-user license and should contain a CrossMark logo, the end user license, and a DOI link to the formal publication on ScienceDirect.

Please refer to Elsevier's posting policy for further information.

18. **For book authors** the following clauses are applicable in addition to the above: Authors are permitted to place a brief summary of their work online only. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version. **Posting to a repository:** Authors are permitted to post a summary of their chapter only in their institution's repository.

19. **Thesis/Dissertation**: If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for Proquest/UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission. Theses and dissertations which contain embedded PJAs as part of the formal submission can be posted publicly by the awarding institution with DOI links back to the formal publications on ScienceDirect.

Elsevier Open Access Terms and Conditions

You can publish open access with Elsevier in hundreds of open access journals or in nearly 2000 established subscription journals that support open access publishing. Permitted third party re-use of these open access articles is defined by the author's choice of Creative Commons user license. See our open access license policy for more information.

Terms & Conditions applicable to all Open Access articles published with Elsevier:

Any reuse of the article must not represent the author as endorsing the adaptation of the article nor should the article be modified in such a way as to damage the author's honour or reputation. If any changes have been made, such changes must be clearly indicated.

The author(s) must be appropriately credited and we ask that you include the end user license and a DOI link to the formal publication on ScienceDirect.

If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source it is the responsibility of the user to ensure their reuse complies with the terms and conditions determined by the rights holder.

Additional Terms & Conditions applicable to each Creative Commons user license:

CC BY: The CC-BY license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article and to make commercial use of the Article (including reuse and/or resale of the Article by commercial entities), provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by/4.0.

CC BY NC SA: The CC BY-NC-SA license allows users to copy, to create extracts, abstracts and new works from the Article, to alter and revise the Article, provided this is not done for commercial purposes, and that the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, indicates if changes were made and the licensor is not represented as endorsing the use made of the work. Further, any new works must be made available on the same conditions. The full details of the license are available at http://creativecommons.org/licenses/by-nc-sa/4.0.

CC BY NC ND: The CC BY-NC-ND license allows users to copy and distribute the Article, provided this is not done for commercial purposes and further does not permit distribution of the Article if it is changed or edited in any way, and provided the user gives appropriate credit (with a link to the formal publication through the relevant DOI), provides a link to the license, and that the licensor is not represented as endorsing the use made of the work. The full details of the license are available at http://creativecommons.org/licenses/by-nc-nd/4.0. Any commercial reuse of Open Access articles published with a CC BY NC SA or CC BY NC ND license requires permission from Elsevier and will be subject to a fee.

Commercial reuse includes:

- Associating advertising with the full text of the Article
- Charging fees for document delivery or access
- Article aggregation
- Systematic distribution via e-mail lists or share buttons

Posting or linking by commercial companies for use by customers of those companies.

20. **Other Conditions**:

v1.10

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Copyright permission for Fig. 1.3:

Cancer Biology & Medicine

Cancer Biology & Medicine Permission/License Terms & Conditions

Licensed content title	Colorectal cancer carcinogenesis: a review of mechanisms
Licensed content author	Kanwal Tariq, Kulsoom Ghias
Licensed content volume (issue)	2016 March;13(1):120-35
License date	October 9, 2021

Please note the following points:

- Permission is granted for your request in both print and electronic formats.
- This permission granted only for the use specified in your request just this time.
- Permission must be acknowledged next to the figure, table, or illustration in print or electronic materials with the format of "Reprinted by permission from Cancer Biology & Medicine. Copyright (YEAR) by Cancer Biology & Medicine".
- There is no charge for your order except for the reuse of the portions in commercial conditions.

Contact Cancer Biology & Medicine by:

Add: Huan-hu-xi Road, He-xi District, Tianjin 300060, China

Phn: 86-22-23522919

Fax: 86-22-23522919

E-mail: editor@cancerbiomed.org

Website: www.cancerbiomed.org

Online submission: https://mc03.manuscriptcentral.com/cbm

(Note that the upper copyright permission [from Cancer Biology & Medicine] was attached to the following email)

Wednesday, November 3, 2021 at 21:26:55 Newfoundland Daylight Time

Subject: Re:Copyright permission of a figure
Date: Saturday, October 9, 2021 at 5:25:08 AM Newfoundland Daylight Time
From: editor
To: Yu, Yajun
Attachments: New-CBM Permission 11.pdf

Dear Dr. Yu,
Thank you for your letter. We are pleased to hear that the figure from an article published in our journal could be reused in your thesis.
Please check the enclosed permission. Should you have further questions please feel free to let me know.
With my best regards,
Ying Li
Scientific Editor, Cancer Biology & Medicine

From: "Yu, Yajun" <yy6084@mun.ca>
Date: 2021-09-24 01:42:41
To: editor@cancerbiomed.org
Subject: Copyright permission of a figure

> Hi Cancer Biology & Medicine,
>
> This is Yajun Yu, a PhD student at the Memorial University of Newfoundland, Canada. I am sending this email to you to kindly ask a question relates to the copyright permission of a figure from a paper published in your journal.
>
> I am writing my thesis in recent days and would like to use Figure 1 of the paper " Tariq K, Ghias K. Colorectal cancer carcinogenesis: a review of mechanisms. Cancer Biol Med. 2016 Mar;13(1):120-35. " in my introduction part. Just want to ask how can I obtain the copyright permission for using that figure? It seems there is no link on the web page of the article that leads to the copyright permission request. Thanks.
>
> Best regards,
>
> Yajun

Page 1 of 2

Copyright permission for Table 1.1:

BMJ PUBLISHING GROUP LTD. LICENSE
TERMS AND CONDITIONS

Sep 23, 2021

This Agreement between Mr. Yajun Yu ("You") and BMJ Publishing Group Ltd. ("BMJ Publishing Group Ltd.") consists of your license details and the terms and conditions provided by BMJ Publishing Group Ltd. and Copyright Clearance Center.

License Number	5154570966467
License date	Sep 23, 2021
Licensed Content Publisher	BMJ Publishing Group Ltd.
Licensed Content Publication	Gut
Licensed Content Title	Genetic architecture of colorectal cancer
Licensed Content Author	Ulrike Peters,Stephanie Bien,Niha Zubair
Licensed Content Date	Oct 1, 2015
Licensed Content Volume	64
Licensed Content Issue	10
Type of Use	Dissertation/Thesis

Requestor type	Individual
Format	Print and electronic
Portion	Figure/table/extract
Number of figure/table/extracts	1
Descriptionof figure/table/extracts	Table 1
Will you be translating?	No
Circulation/distribution	20
Title	Genetic and clinico-demographic factors with and without time-varying associations with survival outcomes in colorectal cancer
Institution name	Memorial University of Newfoundland
Expected presentation date	Nov 2021
Portions	Table 1
Requestor Location	Mr. Yajun Yu Craig L. Dobbin Genetics Research Centre St. John's, NL A1B 3V6 Canada Attn: Mr. Yajun Yu

Publisher Tax ID	GB674738491
Total	0.00 CAD

Terms and Conditions

BMJ Terms and Conditions for Permissions

When you submit your order you are subject to the terms and conditions set out below. You will also have agreed to the Copyright Clearance Center's ("CCC") terms and conditions regarding billing and payment https://s100.copyright.com/App/PaymentTermsAndConditions.jsp. CCC are acting as BMJ Publishing Group Limited's ("BMJs") agent.

Subject to the terms set out herein, BMJ hereby grants to you (the Licensee) a non-exclusive, non-transferable licence to re-use material as detailed in your request for this/those purpose(s) only and in accordance with the following conditions:

1) **Scope of Licence:** Use of the Licensed Material(s) is restricted to the ways specified by you during the order process and any additional use(s) outside of those specified in that request, require a further grant of permission.

2) **Acknowledgement:** In all cases, due acknowledgement to the original publication with permission from BMJ should be stated adjacent to the reproduced Licensed Material. The format of such acknowledgement should read as follows:
"Reproduced from [publication title, author(s), volume number, page numbers, copyright notice year] with permission from BMJ Publishing Group Ltd."

3) **Third Party Material**: BMJ acknowledges to the best of its knowledge, it has the rights to licence your reuse of the Licensed Material, subject always to the caveat that images/diagrams, tables and other illustrative material included within, which have a separate copyright notice, are presumed as excluded from the licence. Therefore, you should ensure that the Licensed Material you are requesting is original to BMJ and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested in any way indicates that it was reprinted or adapted by BMJ with permission from another source, then you should seek permission from that source directly to re-use the Licensed Material, as this is outside of the licence granted herein.

4) **Altering/Modifying Material:** The text of any material for which a licence is granted may not be altered in any way without the prior express permission of BMJ. If adaptation of the material has been approved via bmj.permissions@bmj.com you must include the disclaimer: "Adapted by permission from BMJ Publishing Group Limited. [publication title, author, volume number, page numbers, copyright notice year]

5) **Reservation of Rights:** BMJ reserves all rights not specifically granted in the combination of (i) the licence details provided by you and accepted in the course of this

licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment Terms and Conditions.

6) **Timing of Use:** First use of the Licensed Material must take place within 12 months of the grant of permission.

7) **Creation of Contract and Termination:** Once you have submitted an order via RightsLink and this is received by CCC, and subject to you completing accurate details of your proposed use, this is when a binding contract is in effect and our acceptance occurs. As you are ordering rights from a periodical, to the fullest extent permitted by law, you will have no right to cancel the contract from this point other than for BMJ's material breach or fraudulent misrepresentation or as otherwise permitted under a statutory right. Payment must be made in accordance with CCC's Billing and Payment Terms and conditions. In the event that you breach any material condition of these terms and condition or any of CCC's Billing and Payment Terms and Conditions, the license is automatically terminated upon written notice from BMJ or CCC or as otherwise provided for in CCC's Billing and Payment Terms and Conditions, where these apply. Continued use of materials where a licence has been terminated, as well as any use of the Licensed Materials beyond the scope of an unrevoked licence, may constitute intellectual property rights infringement and BMJ reserves the right to take any and all action to protect its intellectual property rights in the Licensed Materials.

8) **Warranties:** BMJ makes no express or implied representations or warranties with respect to the Licensed Material and to the fullest extent permitted by law this is provided on an "as is" basis. For the avoidance of doubt BMJ does not warrant that the Licensed Material is accurate or fit for any particular purpose.

9) **Limitation of Liability:** To the fullest extent permitted by law, BMJ disclaims all liability for any indirect, consequential or incidental damages (including without limitation, damages for loss of profits, information or interruption) arising out of the use or inability to use the Licensed Material or the inability to obtain additional rights to use the Licensed Material. To the fullest extent permitted by law, the maximum aggregate liability of BMJ for any claims, costs, proceedings and demands for direct losses caused by BMJ's breaches of its obligations herein shall be limited to twice the amount paid by you to CCC for the licence granted herein.

10) **Indemnity:** You hereby indemnify and hold harmless BMJ and their respective officers, directors, employees and agents, from and against any and all claims, costs, proceeding or demands arising out of your unauthorised use of the Licensed Material.

11) **No Transfer of License:** This licence is personal to you, and may not be assigned or transferred by you without prior written consent from BMJ or its authorised agent(s). BMJ may assign or transfer any of its rights and obligations under this Agreement, upon written notice to you.

12) **No Amendment Except in Writing:** This licence may not be amended except in a writing signed by both parties (or, in the case of BMJ, by CCC on BMJ's behalf).

13) **Objection to Contrary terms:** BMJ hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment Terms and Conditions. These terms and conditions, together with CCC's Billing and Payment Terms and Conditions (which to the extent they are consistent are incorporated herein), comprise the entire agreement between you and BMJ (and CCC) and the Licensee

concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment Terms and Conditions, these terms and conditions shall control.

14) **Revocation:** BMJ or CCC may, within 30 days of issuance of this licence, deny the permissions described in this licence at their sole discretion, for any reason or no reason, with a full refund payable to you should you have not been able to exercise your rights in full. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice from BMJ or CCC will not, to the fullest extent permitted by law alter or invalidate the denial. For the fullest extent permitted by law in no event will BMJ or CCC be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to BMJ and/or CCC for denied permissions.

15) **Restrictions to the license**:

15.1) **Promotion:** BMJ will not give permission to reproduce in full or in part any Licensed Material for use in the promotion of the following:

a) non-medical products that are harmful or potentially harmful to health

b) medical products that do not have a product license granted by the Medicines and Healthcare products Regulatory Agency (MHRA) or its international equivalents. Marketing of the product may start only after data sheets have been released to members of the medical profession and must conform to the marketing authorization contained in the product license.

16) **Translation**: This permission is granted for non-exclusive world English language rights only unless explicitly stated in your licence. If translation rights are granted, a professional translator should be employed and it must be a true reproduction, accurately conveying the original meaning and of the same quality.

17) **STM Permissions Guidelines**: For content reuse in journals that qualify for permission under the STM Permissions Guidelines (which may be updated from time to time) the terms and conditions of the Guidelines supersede those in this licence. https://www.stm-assoc.org/intellectual-property/permissions/permissions-guidelines/

18) **General:** Neither party shall be liable for failure, default or delay in performing its obligations under this Licence, caused by a Force Majeure event which shall include any act of God, war, or threatened war, act or threatened act of terrorism, riot, strike, lockout, individual action, fire, flood, drought, tempest or other event beyond the reasonable control of either party.

18.1) In the event that any provision of this Agreement is held to be invalid, the remainder of the provisions shall continue in full force and effect.

18.2) There shall be no right whatsoever for any third party to enforce the terms and conditions of this Agreement. The Parties hereby expressly wish to exclude the operation of the Contracts (Rights of Third Parties) Act 1999 and any other legislation which has this effect and is binding on this agreement.

18.3) To the fullest extent permitted by law, this Licence will be governed by the laws of England and shall be governed and construed in accordance with the laws of England. Any action arising out of or relating to this agreement shall be brought in courts situated in England save where it is necessary for BMJ for enforcement to bring proceedings to

bring an action in an alternative jurisdiction.

V1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Copyright permission for Table 1.2:

SPRINGER NATURE LICENSE
TERMS AND CONDITIONS

Sep 23, 2021

This Agreement between Mr. Yajun Yu ("You") and Springer Nature ("Springer Nature") consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number	5154901149423
License date	Sep 23, 2021
Licensed Content Publisher	Springer Nature
Licensed Content Publication	Nature Reviews Clinical Oncology
Licensed Content Title	Surveillance after curative treatment for colorectal cancer
Licensed Content Author	Eric P. van der Stok et al
Licensed Content Date	Dec 20, 2016
Type of Use	Thesis/Dissertation
Requestor type	non-commercial (non-profit)
Format	print and electronic
Portion	figures/tables/illustrations
Number of figures/tables/illustrations	1

High-res required	no
Will you be translating?	no
Circulation/distribution	1 - 29
Author of this Springer Nature content	no
Title	Genetic and clinico-demographic factors with and without time-varying associations with survival outcomes in colorectal cancer
Institution name	Memorial University of Newfoundland
Expected presentation date	Nov 2021
Portions	Table 2 (note that the format and citation numbers of cited references in the table will be updated to fit the citation system of my thesis)
Requestor Location	Mr. Yajun Yu Craig L. Dobbin Genetics Research Centre St. John's, NL A1B 3V6 Canada Attn: Mr. Yajun Yu
Total	0.00 CAD

Terms and Conditions

Springer Nature Customer Service Centre GmbH
Terms and Conditions

This agreement sets out the terms and conditions of the licence (the **Licence**) between you and **Springer Nature Customer Service Centre GmbH** (the **Licensor**). By clicking 'accept' and completing the transaction for the material (**Licensed Material**), you also confirm your acceptance of these terms and conditions.

1. Grant of License

1. 1. The Licensor grants you a personal, non-exclusive, non-transferable, world-wide licence to reproduce the Licensed Material for the purpose specified in your order only. Licences are granted for the specific use requested in the order and for no other use, subject to the conditions below.

1. 2. The Licensor warrants that it has, to the best of its knowledge, the rights to license reuse of the Licensed Material. However, you should ensure that the material you are requesting is original to the Licensor and does not carry the copyright of another entity (as credited in the published version).

1. 3. If the credit line on any part of the material you have requested indicates that it was reprinted or adapted with permission from another source, then you should also seek permission from that source to reuse the material.

2. Scope of Licence

2. 1. You may only use the Licensed Content in the manner and to the extent permitted by these Ts&Cs and any applicable laws.

2. 2. A separate licence may be required for any additional use of the Licensed Material, e.g. where a licence has been purchased for print only use, separate permission must be obtained for electronic re-use. Similarly, a licence is only valid in the language selected and does not apply for editions in other languages unless additional translation rights have been granted separately in the licence. Any content owned by third parties are expressly excluded from the licence.

2. 3. Similarly, rights for additional components such as custom editions and derivatives require additional permission and may be subject to an additional fee. Please apply to Journalpermissions@springernature.com/bookpermissions@springernature.com for these rights.

2. 4. Where permission has been granted **free of charge** for material in print, permission may also be granted for any electronic version of that work, provided that the material is incidental to your work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.

2. 5. An alternative scope of licence may apply to signatories of the STM Permissions Guidelines, as amended from time to time.

3. Duration of Licence

3. 1. A licence for is valid from the date of purchase ('Licence Date') at the end of the relevant period in the below table:

Scope of Licence	Duration of Licence
Post on a website	12 months
Presentations	12 months
Books and journals	Lifetime of the edition in the language purchased

4. Acknowledgement

4. 1. The Licensor's permission must be acknowledged next to the Licenced Material in print. In electronic form, this acknowledgement must be visible at the same time as the figures/tables/illustrations or abstract, and must be hyperlinked to the journal/book's homepage. Our required acknowledgement format is in the Appendix below.

5. Restrictions on use

5. 1. Use of the Licensed Material may be permitted for incidental promotional use and minor editing privileges e.g. minor adaptations of single figures, changes of format, colour and/or style where the adaptation is credited as set out in Appendix 1 below. Any other changes including but not limited to, cropping, adapting, omitting material that affect the meaning, intention or moral rights of the author are strictly prohibited.

5. 2. You must not use any Licensed Material as part of any design or trademark.

5. 3. Licensed Material may be used in Open Access Publications (OAP) before publication by Springer Nature, but any Licensed Material must be removed from OAP sites prior to final publication.

6. Ownership of Rights

6. 1. Licensed Material remains the property of either Licensor or the relevant third party and any rights not explicitly granted herein are expressly reserved.

7. Warranty

IN NO EVENT SHALL LICENSOR BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL OR INDIRECT DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND
WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.

8. Limitations

8. 1. ***BOOKS ONLY:*** Where **'reuse in a dissertation/thesis'** has been selected the following terms apply: Print rights of the final author's accepted manuscript (for clarity,

NOT the published version) for up to 100 copies, electronic rights for use only on a personal website or institutional repository as defined by the Sherpa guideline (www.sherpa.ac.uk/romeo/).

8. 2. For content reuse requests that qualify for permission under the STM Permissions Guidelines, which may be updated from time to time, the STM Permissions Guidelines supersede the terms and conditions contained in this licence.

9. Termination and Cancellation

9. 1. Licences will expire after the period shown in Clause 3 (above).

9. 2. Licensee reserves the right to terminate the Licence in the event that payment is not received in full or if there has been a breach of this agreement by you.

Appendix 1 — Acknowledgements:

For Journal Content:
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication papers:**
Reprinted by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].)

For Adaptations/Translations:
Adapted/Translated by permission from [**the Licensor**]: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

Note: For any republication from the British Journal of Cancer, the following credit line style applies:

Reprinted/adapted/translated by permission from [**the Licensor**]: on behalf of Cancer Research UK: : [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication)

For **Advance Online Publication** papers:
Reprinted by permission from The [**the Licensor**]: on behalf of Cancer Research UK: [**Journal Publisher** (e.g. Nature/Springer/Palgrave)] [**JOURNAL NAME**] [**REFERENCE CITATION** (Article name, Author(s) Name), [**COPYRIGHT**] (year of publication), advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM])

For Book content:
Reprinted/adapted by permission from [**the Licensor**]: [**Book Publisher** (e.g.

Palgrave Macmillan, Springer etc) [**Book Title**] by [**Book author**(s)]
[**COPYRIGHT**] (year of publication)

Other Conditions:

Version 1.3

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix B: Supporting information for "Two functional indel polymorphisms in the promoter region of the Brahma gene (*BRM*) and disease risk and progression-free survival in colorectal cancer" (Chapter 2)

Supplementary Table 1. Definition of genotype combination categories.

Categories	Genotype combination [*BRM*-741 + *BRM*-1321]	References
Category A.		
Double wild-type genotype	[Del/Del + Del/Del]	[1–5]
No homozygous variant genotype	[Ins/Del + Del/Del] or [Del/Del + Ins/Del] or [Ins/Del + Ins/Del]	
One homozygous variant genotype	[Ins/Ins + Del/Del] or [Ins/Ins + Ins/Del] or [Del/Del + Ins/Ins] or [Ins/Del + Ins/Ins]	
Double homozygous variant genotype	[Ins/Ins + Ins/Ins]	
Category B.		
Double homozygous variant genotype	[Ins/Ins + Ins/Ins]	
Others	[Del/Del + Del/Del] or [Ins/Del + Del/Del] or [Del/Del + Ins/Del] or [Ins/Del + Ins/Del] or [Ins/Ins + Del/Del] or [Ins/Ins + Ins/Del] or [Del/Del + Ins/Ins] or [Ins/Del + Ins/Ins]	
Category C.		
Double wild-type genotype	[Del/Del + Del/Del]	
Others	[Ins/Del + Del/Del] or [Del/Del + Ins/Del] or [Ins/Del + Ins/Del] or [Ins/Ins + Del/Del] or [Ins/Ins + Ins/Del] or [Del/Del + Ins/Ins] or [Ins/Del + Ins/Ins] or [Ins/Ins + Ins/Ins]	
Category D.		
At least one homozygous variant genotype	[Ins/Ins + Del/Del] or [Ins/Ins + Ins/Del] or [Del/Del + Ins/Ins] or [Ins/Del + Ins/Ins] or [Ins/Ins + Ins/Ins]	
Others	[Del/Del + Del/Del] or [Ins/Del + Del/Del] or [Del/Del + Ins/Del] or [Ins/Del + Ins/Del]	

Supplementary Table 2. Results of the multivariable logistic regression analyses for colon and rectal cancer patients.

A. Colon cases (n=280) + controls (n=408)					
			95% CI		
Variables	**Category**	*** OR**	**lower**	**higher**	**p value**
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	1.39	0.95	2.05	0.10
	2 vs 0	1.23	0.78	1.96	0.38
BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	1.35	0.94	1.95	0.11
	2 vs 0	1.12	0.71	1.77	0.61
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	1.65	1.05	2.63	**0.03**
	2 vs 0	1.77	1.05	3.01	**0.03**
	3 vs 0	1.14	0.60	2.15	0.69
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	0.74	0.43	1.25	0.27
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	1.60	1.04	2.51	**0.03**
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	1.05	0.75	1.47	0.78
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	1.34	0.93	1.94	0.12
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.99	0.67	1.44	0.95
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	1.12	0.89	1.40	0.35
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	1.28	0.91	1.81	0.16
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.94	0.63	1.38	0.74
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	1.09	0.87	1.36	0.47
B. Rectum cases (n=146) + controls (n=408)					
			95% CI		
Variables	**Category**	*** OR**	**lower**	**higher**	**p value**
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.73	0.47	1.14	0.16
	2 vs 0	0.62	0.35	1.08	0.09

BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.95	0.62	1.48	0.83
	2 vs 0	0.62	0.34	1.11	0.11
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	0.99	0.60	1.64	0.95
	2 vs 0	0.72	0.38	1.36	0.32
	3 vs 0	0.60	0.27	1.30	0.21
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	0.65	0.31	1.26	0.23
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	0.86	0.54	1.40	0.55
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	0.69	0.44	1.07	0.10
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.69	0.46	1.06	0.09
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.75	0.45	1.22	0.25
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.78	0.59	1.03	0.08
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.85	0.56	1.28	0.43
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.64	0.36	1.07	0.10
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.82	0.62	1.08	0.15

CI, confidence interval; Del, deletion; Ins, insertion; OR, odds ratio. P values < 0.05 are shown in bold.

[*] Adjusted for age, sex, number of first degree relatives with colorectal cancer, smoking status, and body mass index.

Supplementary Table 3. Results of the multivariable logistic regression analyses for male and female patients.

A. Male cases (n=255) + male controls (n=242)					
			95% CI		
Variables	**Category**	*** OR**	**lower**	**higher**	**p value**
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	1.00	0.65	1.53	0.99
	2 vs 0	1.07	0.64	1.79	0.79
BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	1.14	0.76	1.73	0.52
	2 vs 0	1.10	0.66	1.85	0.71
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	1.17	0.72	1.91	0.52
	2 vs 0	1.42	0.81	2.52	0.22
	3 vs 0	1.02	0.51	2.05	0.95
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	0.86	0.47	1.56	0.62
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	1.21	0.77	1.91	0.41
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	1.14	0.77	1.68	0.51
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	1.02	0.68	1.53	0.92
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	1.07	0.70	1.66	0.75
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	1.03	0.80	1.33	0.80
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	1.13	0.77	1.67	0.53
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	1.02	0.65	1.60	0.94
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	1.06	0.82	1.37	0.65
B. Female cases (n=172) + female controls (n=166)					
			95% CI		
Variables	**Category**	*** OR**	**lower**	**higher**	**p value**
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	1.14	0.67	1.97	0.62
	2 vs 0	0.75	0.38	1.46	0.40

BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	1.29	0.78	2.16	0.32
	2 vs 0	0.69	0.35	1.34	0.28
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	1.66	0.87	3.20	0.12
	2 vs 0	1.16	0.54	2.52	0.70
	3 vs 0	0.69	0.26	1.75	0.44
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	0.49	0.22	1.06	0.08
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	1.38	0.75	2.59	0.31
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	0.66	0.40	1.09	0.11
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	1.01	0.61	1.70	0.96
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.69	0.39	1.20	0.19
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.88	0.63	1.22	0.45
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	1.09	0.67	1.77	0.72
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.59	0.33	1.07	0.08
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.89	0.64	1.22	0.46

CI, confidence interval; Del, deletion; Ins, insertion; OR, odds ratio.

[*] Adjusted for age, number of first degree relatives with colorectal cancer, smoking status, and body mass index.

Supplementary Table 4. Results of the age-stratified multivariable Cox regression (survival) analyses in the male and female sub-cohorts.

A. Male cases (n=255)						
			95% CI			
Variables	**Category**	*** HR**	**lower**	**higher**	**p value**	**p value for PH assumption test**
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.54	0.34	0.88	**0.01**	0.45
	2 vs 0	1.29	0.77	2.17	0.33	0.70
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.73	0.48	1.13	0.16	0.67
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	1.84	1.17	2.90	**0.01**	0.89
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	1.09	0.81	1.47	0.56	0.76
BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.96	0.61	1.51	0.86	0.05
	2 vs 0	1.01	0.57	1.78	0.97	0.25
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.97	0.64	1.48	0.90	0.05
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	1.03	0.62	1.72	0.90	0.77
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	1.00	0.75	1.32	0.99	0.14
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	0.64	0.38	1.09	0.10	0.99
	2 vs 0	0.94	0.52	1.69	0.84	0.61
	3 vs 0	1.23	0.61	2.51	0.56	0.94
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	1.56	0.84	2.88	0.16	0.81
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	0.78	0.49	1.26	0.31	0.78
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	1.37	0.9	2.09	0.14	0.71
B. Female cases (n=171)						
			95% CI			
Variables	**Category**	*** HR**	**lower**	**higher**	**p value**	**p value for PH assumption test**
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.96	0.50	1.84	0.89	0.95

	2 vs 0	0.56	0.22	1.39	0.21	0.13
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.83	0.44	1.54	0.55	0.60
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.57	0.25	1.30	0.18	0.09
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.78	0.52	1.18	0.24	0.23
BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.83	0.44	1.59	0.58	0.21
	2 vs 0	0.85	0.34	2.11	0.72	1.00
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.84	0.45	1.55	0.57	0.30
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	0.95	0.42	2.16	0.90	0.54
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.9	0.57	1.41	0.64	0.68
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	0.65	0.30	1.39	0.27	0.24
	2 vs 0	0.45	0.17	1.18	0.10	0.49
	3 vs 0	0.61	0.16	2.31	0.47	0.94
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	0.94	0.29	3.07	0.91	0.61
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	0.59	0.28	1.21	0.15	0.56
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	0.68	0.34	1.35	0.27	0.08

CI, confidence interval; Del, deletion; HR, hazard ratio; Ins, insertion; PH, proportional hazard. P value < 0.05 are shown in bold. P-values are rounded to two decimals.

[*] Age-stratified Cox models adjusted for disease stage, tumor location, microsatellite instability (MSI) status, and treatment with adjuvant chemotherapy status.

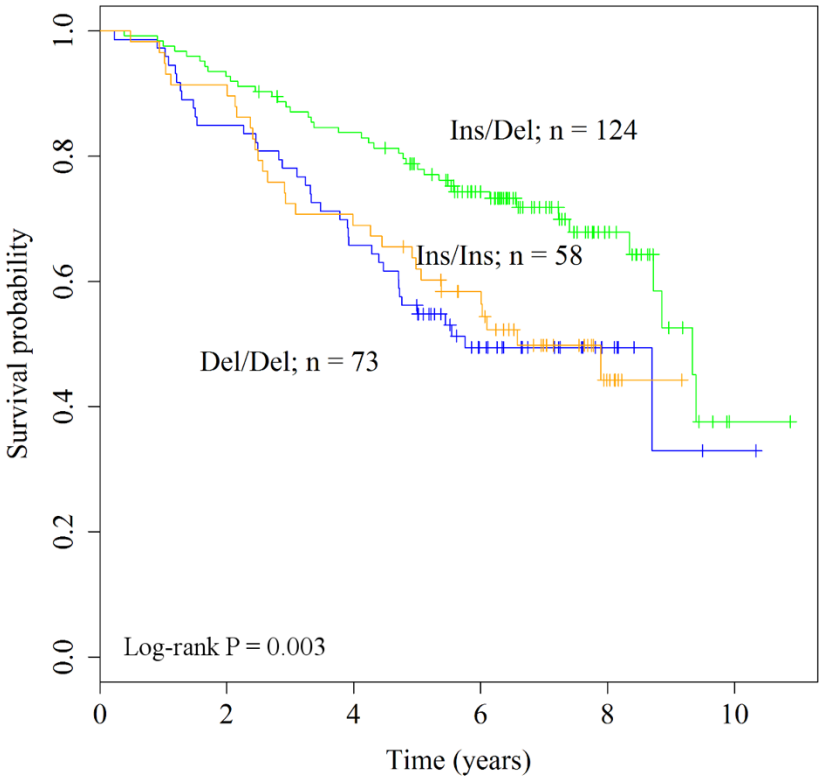
Supplementary Table 5. Results of the age-stratified multivariable Cox regression (survival) analyses in the colon and rectal cancer sub-cohorts.

A. Colon cases (n=280)						
			95% CI			
Variables	**Category**	*** HR**	**lower**	**higher**	**p value**	**p value for PH assumption test**
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.76	0.46	1.26	0.28	0.42
	2 vs 0	0.97	0.53	1.77	0.91	0.43
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.81	0.50	1.31	0.39	0.37
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	1.17	0.71	1.93	0.54	0.71
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.97	0.71	1.34	0.87	0.42
BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	1.00	0.62	1.62	1.00	0.44
	2 vs 0	1.05	0.56	1.97	0.87	0.36
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	1.01	0.64	1.61	0.96	0.36
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	1.05	0.61	1.82	0.85	0.54
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	1.02	0.75	1.40	0.89	0.34
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	0.63	0.35	1.14	0.13	0.63
	2 vs 0	0.69	0.35	1.35	0.28	0.26
	3 vs 0	0.93	0.40	2.16	0.87	0.61
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	1.34	0.67	2.70	0.41	0.47
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	0.67	0.39	1.18	0.16	0.98
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	1.07	0.68	1.68	0.77	0.19
B. Rectum cases (n=146)						
			95% CI			

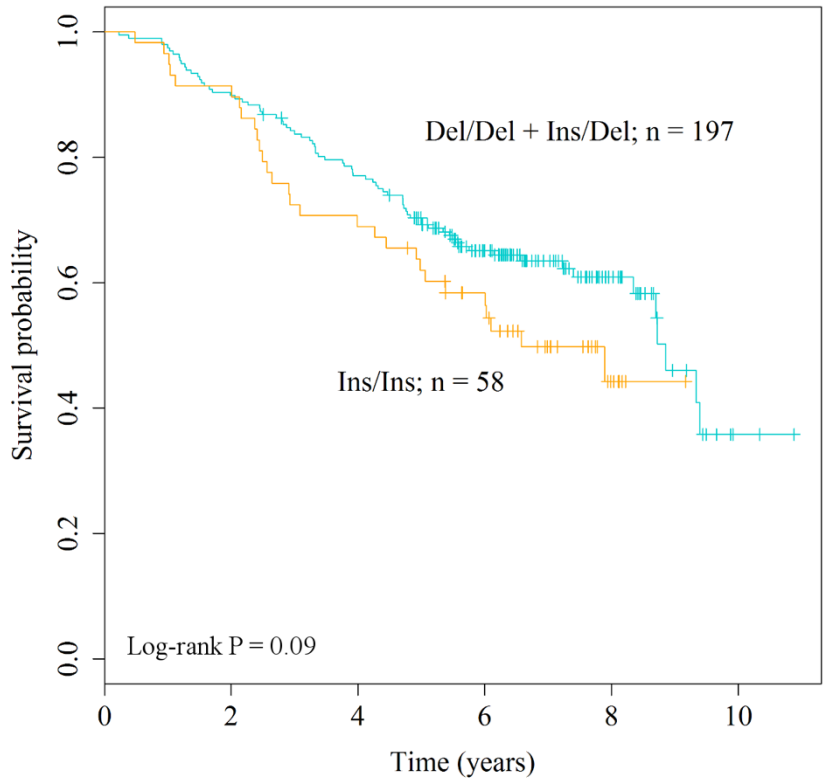
Variables	Category	* HR	lower	higher	p value	p value for PH assumption test
BRM-741 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.59	0.32	1.07	0.08	0.18
	2 vs 0	1.00	0.52	1.93	1.00	0.72
BRM-741 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.71	0.42	1.21	0.21	0.28
BRM-741 (recessive model; 0=Others; 1=Ins/Ins)	1 vs 0	1.29	0.71	2.37	0.40	0.84
BRM-741 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	2 vs 1 vs 0	0.94	0.65	1.34	0.71	0.55
BRM-1321 (co-dominant model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.90	0.51	1.60	0.73	0.90
	2 vs 0	1.02	0.48	2.17	0.95	0.67
BRM-1321 (dominant model; 0=Del/Del; 1=Others)	1 vs 0	0.93	0.55	1.59	0.80	0.95
BRM-1321 (recessive model; 0=Others; 1=Ins/Ins)	2 vs 1 vs 0	1.08	0.54	2.16	0.83	0.61
BRM-1321 (additive model; 0=Del/Del; 1=Ins/Del; 2=Ins/Ins)	1 vs 0	0.99	0.68	1.43	0.95	0.76
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=No Ins/Ins; 2=One Ins/Ins; 3=Both Ins/Ins)	1 vs 0	0.64	0.34	1.19	0.16	0.89
	2 vs 0	0.78	0.37	1.63	0.50	0.61
	3 vs 0	1.08	0.42	2.83	0.87	0.99
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=Both Ins/Ins)	1 vs 0	1.44	0.60	3.46	0.41	0.88
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Both Del/Del; 1=Others)	1 vs 0	0.72	0.41	1.27	0.25	0.78
Genotype combination of *BRM*-741 and *BRM*-1321 (0=Others; 1=At least one Ins/Ins)	1 vs 0	1.14	0.65	1.98	0.66	0.74

CI, confidence interval; Del, deletion; HR, hazard ratio; Ins, insertion; PH, proportional hazard.

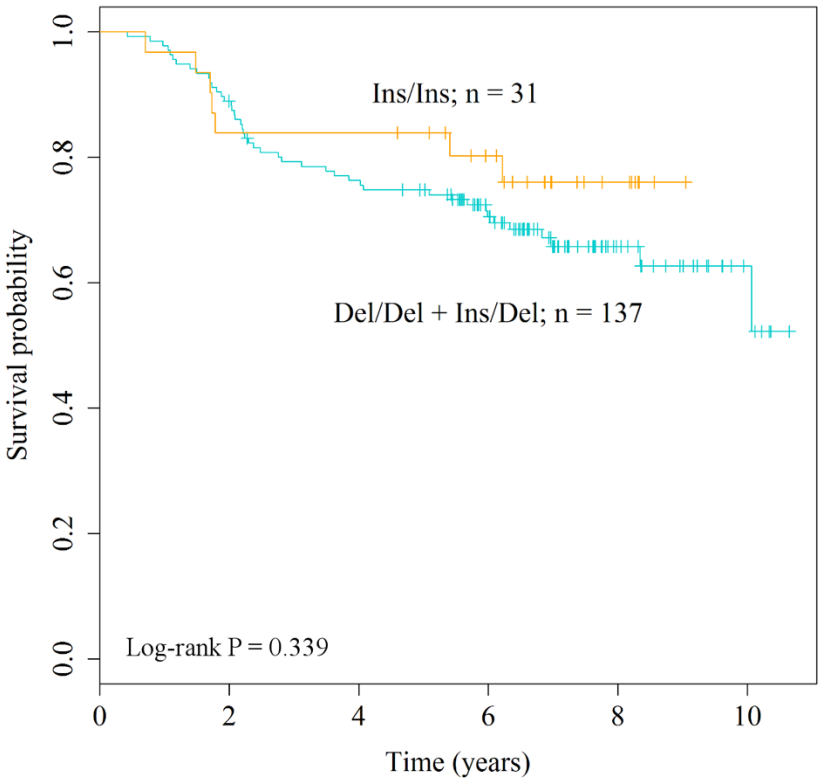
* Age-stratified Cox models adjusted for disease stage, microsatellite instability (MSI) status, and treatment with adjuvant chemotherapy status. Please note that in rectum cases, the MSI was not included as a covariate because there were only two patients with the microsatellite instability-high (MSI-H) tumor type.



Supplementary Figure 1. Kaplan-Meier curves for the *BRM*-741 indel under the co-dominant genetic model in the male colorectal cancer cases.



Supplementary Figure 2. Kaplan-Meier curves for the *BRM*-741 indel under the recessive genetic model in the male colorectal cancer cases.



Supplementary Figure 3. Kaplan-Meier curves for the *BRM*-741 indel under the recessive genetic model in the female colorectal cancer cases.

Appendix C: Supporting information for "Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying associations with the risk of relapse" (Chapter 3)

Detection of INDELs/CNVs.

Since the majority of the recent CNV studies recommended using more than one CNV detection algorithm to increase the prediction accuracy and to decrease the false positive findings [6–8] we used two well-assessed and widely-used CNV calling algorithms; PennCNV [9] and QuantiSNP [10]. Both of these algorithms require signal intensity files of each subject as inputs, which were created using a custom Perl program by merging two types of data files obtained during the genotyping reaction. These data files were; a) the report data files, which included signal intensity data (Log R ratio [LRR] and B allele frequency [BAF] values) obtained during the genotyping reaction for each marker, and b) the final report MAP file that included the chromosome numbers, marker names, and marker positions based on the human genome assembly hg19. The signal intensity files generated for each of the patients were then used by the CNV calling algorithms.

QuantiSNP (version 2) package was downloaded from the QuantiSNP download website (https://sites.google.com/site/quantisnp/downloads) on April 2013 and run using the default parameters [10]. Since the differences in the GC contents among different genomic regions may lead to "genomic waviness" in the signal intensity data and complicate the detection of INDELs/CNVs [11], while running QuantiSNP a GC correction step was also performed.

The PennCNV package was downloaded on May 2013 from the PennCNV website (http://www.openbioinformatics.org/penncnv/). To detect the INDELs/CNVs by the PennCNV algorithm, new Population Frequency of B allele (PFB) and GC-model files were required. This was because the PFB and GC-model files provided by the PennCNV website were based on hg18 whereas our data was based on hg19.

To generate the new PFB file, an Illumina® Human Omni1_QuadV1 dataset containing the signal intensity files for 88 HapMap CEU (Caucasian) individuals was downloaded on May 2013 from the Gene Expression Omnibus (GEO) database [12] (platform number: GPL8882 and series number: GSE17197). These signal intensity files were previously created based on the hg18 genome coordinates and uploaded to the database by Illumina®. Second, the HumanOmni1-Quad v1.0 Build 36 to Build 37 Mapping Information file (also named HumanOmni1-Quad_v1-0_B-H_MappingInformation.txt), which included the hg18 genome coordinate information and their equivalent for the hg19, was downloaded on January 2014 from the Illumina® support website (http://support.illumina.com/downloads/humanomni1-quad_product_support_files.html). This mapping information file was then used to substitute the hg18 genome coordinate information with the hg19 information in the 88 HapMap CEU signal intensity files using custom Perl programs. Finally, the reformatted 88 HapMap CEU signal

intensity files were used to generate the final PFB file using the Perl program *Compile_PFB.pl* provided within the PennCNV package [9].

In order to generate the GC-model file (required to correct for genomic waviness) based on the hg19 genome coordinates, we utilized two data files; the first one was the *GC5Base.txt* file that contained the percentage of the GC bases in 5-base windows based on the hg19 genome coordinates. This file was downloaded on January 2014 from the University of California Santa Cruz (UCSC) [13] genome bioinformatics download website (http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/gc5Base.txt.gz) as suggested by the PennCNV developers. Second, we chose the signal intensity file of a randomly selected patient and used it in the Perl program *Cal_gc_snp.pl* that was provided with the PennCNV package to generate the GC model file based on hg19. After the generation of the required input files (that is, patient signal intensity files, PFB file and GC model file all based on the hg19 genome coordinates), INDELs/CNVs in the patient genomes were predicted by the Perl program *Detect_cnv.pl* of PennCNV using the default parameters.

Since it has been noted earlier [9] that PennCNV tends to split large CNVs into smaller ones when SNP genotype array data are used for variant detection, the adjacent variants detected by PennCNV were merged together following the run. This step was performed using the *Clean_cnv.pl* program of the PennCNV package. During this step, variants were merged if the sequence gap between the variants did not exceed 1/2 of the total distance from the start position of the first variant to the end position of the second variant.

In both the QuantiSNP and PennCNV analyses, similar to many other studies [14,15] the INDEL/CNV predictions were limited to the autosomal chromosomes due to the hemizygosity in sex chromosomes that complicates the variant detection.

Initially, QuantiSNP and PennCNV detected 336,288 and 204,439 INDELs/CNVs in the genomes of the 505 patients, respectively (**Supplementary Text Table 1**). Overall, QuantiSNP detected more variations than PennCNV, both algorithms predicted a higher number of CNVs than INDELs, and deletions constituted the majority of the variants.

Supplementary Text Table 1. The main features of the INDELs/CNVs initially predicted by QuantiSNP and PennCNV.

Number of INDELs/CNVs		**QuantiSNP**		**PennCNV**	
Total predicted INDELs/CNVs in the cohort		336,288		204,439	
Average number of INDELs/CNVs per individual		665.92		404.83	
Type		**N**	**%**	**N**	**%**
INDELs		76,854	22.85	46,616	22.80
CNVs		259,434	77.15	157,823	77.20
INDELs/CNVs per CN state		**N**	**%**	**N**	**%**
(CN= 0)	Two copy deletion	76,035	22.61	57,698	28.22
(CN= 1)	One copy deletion	128,908	38.33	94,917	46.43
(CN= 3)	One copy duplication	64,217	19.1	49,983	24.45
*(CN= 4, 5)	Two or more copy duplication	67,128	19.96	1,841	0.90

N: Number, **CN**: Copy number state. *Please note that QuantiSNP [10] assigns the CN state 4 for variants that exist in 4 copies and CN state 5 for variants that exist in 5 or more copies in a genome. However, PennCNV assigns the CN state 4 for variants that exist in 4 or more copies in a genome [9].

Quality control (QC) analysis and further characterization of INDELs/CNVs

After the initial predictions of the variants, the QC files generated by the QuantiSNP and PennCNV for both the patient data and the variants detected were used to exclude samples and variants with low quality. QC parameters implemented were selected based on other groups' works (**Supplementary Text Table 2**). Perl programming (including custom programs written for specific purposes as well as the Perl program *filter_cnv.pl* provided in the PennCNV website) was used to identify and exclude the data that do not meet the QC thresholds. After the post-detection QC analyses, a total of 85,469 INDELs/CNVs from 501 patients and 159,050 INDELs/CNVs from 497 patients satisfied the QC criteria of QuantiSNP and PennCNV, respectively. The data from 495 out of 505 patients satisfied the QC requirements of both algorithms and were used in the remaining steps of the study.

Supplementary Text Table 2. Exclusion criteria for the subjects and INDELs/CNVs detected.

Exclusion Criteria		QuantiSNP	PennCNV	References
Subject filtering	LRR Standard Deviation (LRR_SD)	> 0.28	> 0.28	[7,16]
	BAF Standard Deviation (BAF_SD)	> 0.20	-	[17]
	BAF drift	-	> 0.01	[16]
	LRR waviness factor (WF)	-	≤ - 0.04 and ≥ 0.04	[18] *
	BAF median	-	< 0.45 or > 0.55	[19,20]
	INDEL/CNV number per sample	> Mean + 3 SD	> Mean + 3 SD	[6,7]
	Samples with extremely long CNVs	> 7.5 Mbps	> 7.5 Mbps	[6,7]
INDEL/CNV filtering	Variant length	< 10 bps	< 10 bps	[21,22]
	Number of probes per INDEL/CNV	< 10 probes	< 10 probes	[23,24]
	Confidence Score	< 30 (Max Log Bayes Factor)	<10	[16,17,20,21,25–27]

BAF: B Allele Frequency; bp: base pair; CNV: Copy Number Variation; INDEL: Insertion/Deletion; LRR: Log R Ratio; SD: Standard Deviation; WF: Waviness Factor. *http://www.openbioinformatics.org/penncnv/.

Further inclusion/exclusion filtering were performed to reduce the methodological artifacts, minimize the false positive findings, and eliminate the low quality data as follows:

a) Plink statistical tool [28] was utilized to check whether either of the algorithms predicted two INDELs/CNVs that were overlapping with each other in the same patient's genome. Such variants would be predicted by mistake twice by an algorithm. Of note, we have detected such variants in neither the QuantiSNP nor the PennCNV outputs.

b) A custom Perl program was written to identify the INDELs/CNVs in the same patient that were predicted by both algorithms with the same copy number state (CN) and had sequences overlapping by at least 50% of their lengths. The overlapped variations (n=74,261) were assumed to be same variants that were detected by both algorithms. Interestingly, a large number of these variants (n=62,567, 84.3%) had identical start and end positions. This indicates that the high concordance of border detection if variants are detected by both PennCNV and QuantiSNP. Variations that did not satisfy these criteria were assumed to be predicted by one algorithm only and removed from the list of variants [7,8]. When the borders of the overlapped variants were different from each other, they were merged together and the new variant borders were determined by the smallest start position and the largest end position of the merged variants [7].

c) It is a common practice to exclude INDELs/CNVs detected within the highly repetitive DNA regions, such as centromere and telomere regions, leukocyte immunoglobulin-like receptor gene cluster, and olfactory receptor (OR) gene regions that can complicate the INDEL/CNV detection [9,14]. Hence, any variant that intersected at least 1 bp with these DNA regions were excluded from further analyses by Perl programming. To do so, first a list of centromere and telomere regions, leukocyte immunoglobulin-like receptor gene cluster and OR gene coordinates was generated based on hg19 following a number of resources. In short; **i)** the genome coordinate information for leukocyte immunoglobulin-like receptor gene cluster based on hg18 was obtained from the PennCNV website (http://penncnv.openbioinformatics.org/en/latest/) on February 2014. Then, the "LiftOver" tool of the UCSC [13] genome browser was then used on February 2014 to change the genome coordinates of the leukocyte immunoglobulin-like receptor gene cluster from hg18 to hg19 genome coordinates; **ii)** the list of centromere positions based on hg19 was obtained from the PennCNV website and were adjusted by adding and subtracting 100 kbps to upstream and downstream of each centromere, respectively, following the PennCNV recommendations; **iii)** the UCSC [13] genome browser was utilized to identify the start and end positions of each chromosome based on hg19. Then the telomere regions were determined by adding and subtracting 500 kbps at the start and end positions of each chromosome respectively, as suggested in the PennCNV package; and finally, **iv)** the hg19 genome coordinates of OR genes (n=840) were downloaded from the Human Olfactory Receptors Data Explorer (HORDE) database (http://genome.weizmann.ac.il/horde/) on February 2014 [29].

After excluding 2,905 INDELs/CNVs that overlapped at least 1 bp with highly repetitive genomic regions, 71,356 variants remained in the dataset.

d) Similar to other studies [20,30], INDELs/CNVs that overlap with the previously reported variants (i.e. variants detected by DNA analyses in other studies) were identified. This step was undertaken to further remove possible false predictions from our results and to identify the variants that were most likely to exist in patient genomes (i.e. not false-positives). The sequence overlap criterion was at least 50%. This reference variant information was based on three studies [31–33] that was available at the Database of Genomic Variants (DGV) [34]. As a result, around 97% of the INDELs/CNVs (n=69,290) were found to overlap at least 50% of their lengths with the previously and experimentally identified variants. These INDELs/CNVs were thus detected with a "high confidence", were highly likely to exist in the patient genomes, and constituted the final list of the INDELs/CNVs of this study (**Table 3.2**).

Within the patient cohort, all high confidence variants with the same borders were considered as the same variant and called as "distinct" variants throughout the manuscript. Also, all distinct variations overlapped at least 1 bp with each other were also clustered in copy number variations regions (CNVRs).

Throughout the study, PLINK [28] was utilized to handle and modify data/files, and to define and describe the INDELs/CNVs based on their lengths, frequency and copy number (CN) states.

Oligonucleotides and amplification conditions for duplex end-point PCR.

The DNA sequence for each INDEL/CNV was obtained from the UCSC genome browser [13]. The repetitive sequences were masked in each DNA sequence using RepeatMasker (http://www.repeatmasker.org) [35]. These sequences were then used to design primer pairs using the Primer 3 tool under the default conditions (http://bioinfo.ut.ee/primer3-0.4.0/) [36]. The primer sequences for amplification of a part of the Albumin gene (which was used as the control amplicon in the duplex reactions) were obtained from a previously published study [37]. **Supplementary Text Table 3** shows the sequences of oligonucleotides, and the size and the genomic coordinates of the amplicons.

Duplex PCRs were performed as follows: 1 µL of genomic DNA (6 ng / µL) was amplified in a 10 µL reaction containing 5 µL of AmpliTaq Gold 360 Master Mix (2X; Applied Biosystems, Foster City, USA), 0.25 µL of 360 GC Enhancer (Applied Biosystems, Foster City, USA), 0.5 µM of each primer (Integrated DNA Technologies, Inc., Coralville, Iowa), and 1.75 µL of sterile water. In cases of poor amplification, PCR was repeated using 12 ng of DNA template. All reactions were carried out in MicroAmp Fast Optical 96-Well Reaction Plates (with barcode, 0.1 mL-Catalogue # 4346906, Applied Biosystems, Foster City, USA) using a Veriti 96-Well Fast Thermal Cycler (Applied Biosystems, Foster City, USA) under the following cycling conditions: 10 minutes at 95°C; 30 cycles of 30 seconds at 95°C, 30 seconds at 57°C, and 60 seconds at 72°C; and finally 7 minutes at 72°C. Non-template controls were included to check for DNA contamination in each reaction mix. In addition to the patient DNA samples, the duplex PCR reactions also included amplification of two commercial DNA samples (Catalogue # G1521 and G1471, Promega, Madison, USA) as controls.

PCR products were analyzed by electrophoresis on 2% - 3% agarose gels stained with SYBR Safe DNA Gel Stain (Invitrogen, Foster City, USA), and visualized using AlphaImager EP (ProteinSimple, San Jose, California).

During this analysis five DNA samples were genotyped twice for each variant; in all cases the genotypes obtained were 100% concordant.

Supplementary Text Table 3. Oligonucleotides used for amplification of selected INDELs/CNVs.

Primer Name	Sequence	Genomic Location of the Amplicon	Amplicon Size (bp)	Concordance Rate
ADAM3A/ADAM5A-F	5'-ATC TCT GGG AAA GCC TGG AT-3'	chr8: 39,310,533 - chr8: 39,310,738	205	98%
ADAM3A/ADAM5A-R	5'-ACT TAG CTG CCA TTC CCT CA-3'			
CNOT1-F	5'-CCATCAAAAGGGCACTGATT-3'	chr16: 58,647,478 - chr16: 58,647,719	241	100%
CNOT1-R	5'-GCGACCAATTTTCTACTTTGA-3'			
DLEU1-F	5'-AGGCTTACTTCCAGGTGCAT-3'	chr13: 51,070,494 - chr13: 51,070,665	171	100%
DLEU1-R	5'-TCACCAAGTGGCTACGATCA-3'			
FAM149A-F	5'-CAG TGG CAA AAT CTC CCA AG-3'	chr4: 187,093,696 - chr4: 187,093,942	246	100%
FAM149A-R	5'-AAG GTG TCA TTG CAG TGG TG-3'			
FILIP1L/CMSS1	5'-TGGTTTGGGACACACTGACT-3'	chr3: 99,629,157 - chr3: 99,629,378	221	100%
FILIP1L/CMSS1	5'-CAACATGCATCTGCCACTTC-3'			
LCE3C-F	5'-AGT TGT CCC TCA CCC AAG TG-3'	chr1: 152,573,251 - chr1: 152,573,402	151	93%
LCE3C-R	5'-ATT GAT GGG ACC TGA AGT GC-3'			
NME7-F	5'-AAA TCC AGC ACA GGG ATC TG-3'	chr1: 169,230,753 - chr1: 169,230,911	158	99%
NME7-R	5'-TGC CAT CAT CAG AGT CAA GC-3'			
REV1-F	5'-TCG TCT CCT GAC TTG CCT TT-3'	chr2: 100,104,021 - chr2: 100,104,211	190	100%
REV1-R	5'-GCA TTG TGG GTC TTT CTG CT-3'			
WDR34/VTI1BP4-F	5'-TGGCTTTCACTTGGCTTTCT-3'	chr9: 131,412,585 - chr9: 131,412,785	200	100%
WDR34/VTI1BP4-F	5'-TTTCTTGCCACGTCCCTATC-3'			
WWOX-F	5'-ATG TCA GTG TCC CCC ACA AT-3'	chr16: 78,381,127 - chr16: 78,381,337	210	100%
WWOX-R	5'-GTC AAG AGT GCT GTG CCA AA-3'			

bp: base pair; chr: chromosome; F: forward; R: reverse. Genomic coordinates are based on hg19.

Supplementary Table 6. Information on the 106 INDELs/CNVs.

Variant type	Length	CHR	Start	End	CN state	CN freq.	Gene	Ensembl ID	% Overlap	Overlap type
INDEL	906	1	17676291	17677196	0	0.109	*PADI4*	ENSG00000159339	100.0	Variant located inside the gene
INDEL	521	1	58744143	58744663	0	0.762	*DAB1*	ENSG00000173406	100.0	Variant located inside the gene
CNV	45280	1	72766413	72811692	0	0.343	*RPL31P12*	ENSG00000227207	0.8	Variant covers the whole gene
CNV	2006	1	89476427	89478432	0	0.103	*GBP3*	ENSG00000117226	100.0	Variant located inside the gene
CNV	1117	1	92232111	92233227	0	0.101	*TGFBR3*	ENSG00000069702	100.0	Variant located inside the gene
CNV	30855	1	152556085	152586939	0	0.331	*LCE3C*	ENSG00000244057	1.4	Variant covers the whole gene
CNV	30855	1	152556085	152586939	0	0.331	*LCE3B*	ENSG00000187238	0.9	Variant covers the whole gene
CNV	30855	1	152556085	152586939	3	0.016	*LCE3C*	ENSG00000244057	1.4	Variant covers the whole gene
CNV	30855	1	152556085	152586939	3	0.016	*LCE3B*	ENSG00000187238	0.9	Variant covers the whole gene
CNV	33950	1	169207360	169241309	0	0.117	*NME7*	ENSG00000143156	100.0	Variant located inside the gene
CNV	33950	1	169207360	169241309	1	0.333	*NME7*	ENSG00000143156	100.0	Variant located inside the gene
INDEL	555	1	179607382	179607936	0	0.129	*TDRD5*	ENSG00000162782	100.0	Variant located inside the gene
INDEL	601	1	207292578	207293178	0	0.453	*C4BPA*	ENSG00000123838	100.0	Variant located inside the gene
CNV	2583	2	33224605	33227187	0	0.101	*LTBP1*	ENSG00000049323	100.0	Variant located inside the gene
CNV	38030	2	34698447	34736476	0	0.147	*AC073218.1*	ENSG00000226785	100.0	Variant located inside the gene
CNV	38030	2	34698447	34736476	1	0.220	*AC073218.1*	ENSG00000226785	100.0	Variant located inside the gene
CNV	1713	2	54565729	54567441	0	0.323	*C2orf73*	ENSG00000177994	100.0	Variant located inside the gene

CNV	1862	2	54565729	54567590	0	0.280	*C2orf73*	ENSG00000177994	100.0	Variant located inside the gene
CNV	1862	2	54565729	54567590	1	0.020	*C2orf73*	ENSG00000177994	100.0	Variant located inside the gene
INDEL	413	2	70125092	70125504	0	0.152	*SNRNP27*	ENSG00000124380	100.0	Variant located inside the gene
INDEL	413	2	70125092	70125504	0	0.152	*MXD1*	ENSG00000059728	100.0	Variant located inside the gene
CNV	1262	2	100103752	100105013	0	0.200	*REV1*	ENSG00000135945	100.0	Variant located inside the gene
CNV	1428	2	159959587	159961014	0	0.305	*TANC1*	ENSG00000115183	100.0	Variant located inside the gene
CNV	1865	2	159959587	159961451	0	0.176	*TANC1*	ENSG00000115183	100.0	Variant located inside the gene
INDEL	540	2	182856938	182857477	0	0.111	*PPP1R1C*	ENSG00000150722	100.0	Variant located inside the gene
CNV	1844	2	215728845	215730688	0	0.137	*AC072062.1*	ENSG00000229267	100.0	Variant located inside the gene
CNV	4671	3	32102055	32106725	0	0.196	*NIFKP7*	ENSG00000251590	27.0	Variant partially overlaps with the gene
CNV	4671	3	32102055	32106725	0	0.196	*OSBPL10*	ENSG00000144645	100.0	Variant located inside the gene
CNV	4671	3	32102055	32106725	1	0.002	*NIFKP7*	ENSG00000251590	27.0	Variant partially overlaps with the gene
CNV	4671	3	32102055	32106725	1	0.002	*OSBPL10*	ENSG00000144645	100.0	Variant located inside the gene
CNV	2627	3	47490712	47493338	0	0.141	*SCAP*	ENSG00000114650	100.0	Variant located inside the gene
INDEL	746	3	99628822	99629567	0	0.103	*CMSS1*	ENSG00000184220	100.0	Variant located inside the gene
INDEL	746	3	99628822	99629567	0	0.103	*FILIP1L*	ENSG00000168386	100.0	Variant located inside the gene
INDEL	746	3	99628822	99629567	1	0.158	*CMSS1*	ENSG00000184220	100.0	Variant located inside the gene
INDEL	746	3	99628822	99629567	1	0.158	*FILIP1L*	ENSG00000168386	100.0	Variant located inside the gene
CNV	2092	3	107038162	107040253	0	0.101	*LINC00883*	ENSG00000243701	100.0	Variant located inside the gene

CNV	2092	3	107038162	107040253	0	0.101	*RP11-446H18.5*	ENSG00000239828	100.0	Variant T located inside the gene
INDEL	541	3	124936371	124936911	0	0.166	*SLC12A8*	ENSG00000221955	100.0	Variant located inside the gene
INDEL	541	3	124936371	124936911	1	0.004	*SLC12A8*	ENSG00000221955	100.0	Variant located inside the gene
CNV	4666	3	131708352	131713017	0	0.123	*CPNE4*	ENSG00000196353	100.0	Variant located inside the gene
CNV	4666	3	131708352	131713017	1	0.044	*CPNE4*	ENSG00000196353	100.0	Variant located inside the gene
CNV	5050	3	136021052	136026101	0	0.129	*PCCB*	ENSG00000114054	100.0	Variant located inside the gene
CNV	5050	3	136021052	136026101	1	0.004	*PCCB*	ENSG00000114054	100.0	Variant located inside the gene
INDEL	554	3	159257057	159257610	0	0.123	*IQCJ-SCHIP1*	ENSG00000250588	100.0	Variant located inside the gene
CNV	3201	3	162765807	162769007	0	0.521	*RP11-10O22.1*	ENSG00000241168	100.0	Variant located inside the gene
CNV	3087	3	189737354	189740440	0	0.101	*LEPREL1*	ENSG00000090530	100.0	Variant located inside the gene
CNV	3087	3	189737354	189740440	1	0.006	*LEPREL1*	ENSG00000090530	100.0	Variant located inside the gene
CNV	9416	3	192875738	192885153	0	0.483	*RP11-143P4.2*	ENSG00000232130	100.0	Variant located inside the gene
CNV	9416	3	192875738	192885153	4	0.002	*RP11-143P4.2*	ENSG00000232130	100.0	Variant located inside the gene
CNV	2737	4	91933043	91935779	0	0.141	*CCSER1*	ENSG00000184305	100.0	Variant located inside the gene
INDEL	533	4	115928747	115929279	0	0.105	*NDST4*	ENSG00000138653	100.0	Variant located inside the gene
INDEL	647	4	138966505	138967151	0	0.133	*LINC00616*	ENSG00000248307	100.0	Variant located inside the gene
CNV	1121	4	146438871	146439991	0	0.164	*SMAD1*	ENSG00000170365	100.0	Variant located inside the gene
CNV	1121	4	146438871	146439991	1	0.022	*SMAD1*	ENSG00000170365	100.0	Variant located inside the gene
CNV	1187	4	166003471	166004657	0	0.776	*TMEM192*	ENSG00000170088	100.0	Variant located inside the gene

CNV	3802	4	172989075	172992876	0	0.216	*GALNTL6*	ENSG00000174473	100.0	Variant located inside the gene
INDEL	501	4	182056607	182057107	0	0.309	*LINC00290*	ENSG00000248197	100.0	Variant located inside the gene
CNV	2092	4	186441932	186444023	0	0.564	*PDLIM3*	ENSG00000154553	100.0	Variant located inside the gene
CNV	2179	4	186441932	186444110	0	0.204	*PDLIM3*	ENSG00000154553	100.0	Variant located inside the gene
CNV	4515	4	187093557	187098071	0	0.313	*FAM149A*	ENSG00000109794	5.8	Variant partially overlaps with the gene
CNV	1915	5	1178511	1180425	0	0.180	*CTD-3080P12.3*	ENSG00000249201	10.9	partially overlaps with the gene
CNV	1648	5	21450792	21452439	0	0.139	*GUSBP1*	ENSG00000183666	100.0	Variant located inside the gene
CNV	1498	5	106324802	106326299	0	0.180	*CTC-254B4.1*	ENSG00000251027	100.0	Variant located inside the gene
CNV	1001	5	147553186	147554186	0	0.327	*SPINK14*	ENSG00000196800	100.0	Variant located inside the gene
CNV	1222	6	666535	667756	0	0.388	*EXOC2*	ENSG00000112685	100.0	Variant located inside the gene
INDEL	501	6	18402172	18402672	0	0.402	*snoU13*	ENSG00000238458	20.2	Variant covers the whole gene
INDEL	501	6	18402172	18402672	0	0.402	*RNF144B*	ENSG00000137393	100.0	Variant located inside the gene
INDEL	568	6	51736175	51736742	0	0.251	*PKHD1*	ENSG00000170927	100.0	Variant located inside the gene
INDEL	568	6	51736175	51736742	1	0.057	*PKHD1*	ENSG00000170927	100.0	Variant located inside the gene
CNV	4098	6	53929777	53933874	0	0.473	*MLIP-AS1*	ENSG00000235050	100.0	Variant located inside the gene
CNV	4098	6	53929777	53933874	0	0.473	*MLIP*	ENSG00000146147	100.0	Variant located inside the gene
CNV	1627	6	65347533	65349159	0	0.289	*EYS*	ENSG00000188107	100.0	Variant located inside the gene
INDEL	390	6	89921782	89922171	0	0.317	*GABRR1*	ENSG00000146276	100.0	Variant located inside the gene
INDEL	880	6	167488211	167489090	0	0.176	*RP11-517H2.6*	ENSG00000272980	100.0	Variant located inside the gene

CNV	1389	7	89810608	89811996	0	0.131	*STEAP2-AS1*	ENSG00000227646	100.0	Variant located inside the gene
CNV	1389	7	89810608	89811996	0	0.131	*STEAP2*	ENSG00000157214	100.0	Variant located inside the gene
CNV	1507	7	89810608	89812114	0	0.135	*STEAP2-AS1*	ENSG00000227646	100.0	Variant located inside the gene
CNV	1507	7	89810608	89812114	0	0.135	*STEAP2*	ENSG00000157214	100.0	Variant located inside the gene
CNV	2798	7	126048572	126051369	0	0.121	*AC000370.2*	ENSG00000241921	100.0	Variant located inside the gene
CNV	2905	7	126048572	126051476	0	0.178	*AC000370.2*	ENSG00000241921	100.0	Variant located inside the gene
CNV	1888	7	148074379	148076266	0	0.675	*CNTNAP2*	ENSG00000174469	100.0	Variant located inside the gene
CNV	4441	8	594761	599201	0	0.677	*ERICH1*	ENSG00000104714	100.0	Variant located inside the gene
CNV	1196	8	4122961	4124156	0	0.426	*CSMD1*	ENSG00000183117	100.0	Variant located inside the gene
CNV	1409	8	11245641	11247049	0	0.176	*C8orf12*	ENSG00000184608	100.0	Variant located inside the gene
CNV	3753	8	25066884	25070636	0	0.596	*DOCK5*	ENSG00000147459	100.0	Variant located inside the gene
CNV	153836	8	39233344	39387179	0	0.172	*ADAM5*	ENSG00000196115	27.0	Variant partially overlaps with the gene
CNV	153836	8	39233344	39387179	0	0.172	*ADAM3A*	ENSG00000197475	46.7	Variant covers the whole gene
CNV	153836	8	39233344	39387179	1	0.343	*ADAM5*	ENSG00000196115	27.0	Variant partially overlaps with the gene
CNV	153836	8	39233344	39387179	1	0.343	*ADAM3A*	ENSG00000197475	46.7	Variant covers the whole gene
CNV	2303	8	75364528	75366830	0	0.349	*GDAP1*	ENSG00000104381	100.0	Variant located inside the gene
CNV	3498	8	137160319	137163816	0	0.265	*RP11-149P24.1*	ENSG00000253248	100.0	Variant located inside the gene
CNV	1884	9	71741217	71743100	0	0.527	*TJP2*	ENSG00000119139	100.0	Variant located inside the gene
CNV	2022	9	101309058	101311079	0	0.240	*GABBR2*	ENSG00000136928	100.0	Variant located inside the gene

CNV	1305	9	131412549	131413853	0	0.323	*VTI1BP4*	ENSG00000227759	44.2	Variant partially overlaps with the gene
CNV	1305	9	131412549	131413853	0	0.323	*WDR34*	ENSG00000119333	100.0	Variant located inside the gene
CNV	1337	9	131412549	131413885	0	0.493	*VTI1BP4*	ENSG00000227759	45.5	Variant partially overlaps with the gene
CNV	1337	9	131412549	131413885	0	0.493	*WDR34*	ENSG00000119333	100.0	Variant located inside the gene
INDEL	969	9	138479177	138480145	0	0.362	*RP11-98L5.4*	ENSG00000224045	30.7	Variant partially overlaps with the gene
CNV	1672	10	4708627	4710298	0	0.533	*LINC00704*	ENSG00000231298	100.0	Variant located inside the gene
CNV	1257	10	27000558	27001814	0	0.352	*PDSS1*	ENSG00000148459	100.0	Variant located inside the gene
CNV	4822	10	78255873	78260694	0	0.648	*C10orf11*	ENSG00000148655	100.0	Variant located inside the gene
INDEL	520	10	89275888	89276407	0	0.430	*MINPP1*	ENSG00000107789	100.0	Variant located inside the gene
INDEL	520	10	89275888	89276407	1	0.002	*MINPP1*	ENSG00000107789	100.0	Variant located inside the gene
INDEL	738	10	95545536	95546273	0	0.891	*LGI1*	ENSG00000108231	100.0	Variant located inside the gene
CNV	2987	10	114113589	114116575	0	0.129	*GUCY2GP*	ENSG00000243316	91.9	Variant partially overlaps with the gene
CNV	1588	10	122226947	122228534	0	0.145	*PPAPDC1A*	ENSG00000203805	100.0	Variant located inside the gene
CNV	2181	11	5760106	5762286	0	0.135	*TRIM5*	ENSG00000132256	100.0	Variant located inside the gene
INDEL	472	11	9324025	9324496	0	0.230	*TMEM41B*	ENSG00000166471	100.0	Variant located inside the gene
CNV	3369	11	31394060	31397428	0	0.133	*DNAJC24*	ENSG00000170946	100.0	Variant located inside the gene
CNV	3369	11	31394060	31397428	1	0.117	*DNAJC24*	ENSG00000170946	100.0	Variant located inside the gene
CNV	1005	11	45430401	45431405	0	0.297	*RP11-430H10.4*	ENSG00000255041	100.0	Variant located inside the gene
INDEL	877	11	66712229	66713105	0	0.232	*PC*	ENSG00000173599	100.0	Variant located inside the gene

INDEL	432	12	12026506	12026937	0	0.121	*ETV6*	ENSG00000139083	100.0	Variant located inside the gene
INDEL	760	12	16420184	16420943	0	0.182	*SLC15A5*	ENSG00000188991	100.0	Variant located inside the gene
CNV	6414	12	45903118	45909531	0	0.289	*RP11-352M15.1*	ENSG00000257657	100.0	Variant located inside the gene
CNV	6414	12	45903118	45909531	1	0.065	*RP11-352M15.1*	ENSG00000257657	100.0	Variant located inside the gene
INDEL	601	13	39934551	39935151	0	0.366	*LHFP*	ENSG00000183722	100.0	Variant located inside the gene
CNV	3249	13	51069352	51072600	0	0.535	*DLEU1*	ENSG00000176124	100.0	Variant located inside the gene
CNV	2194	13	101894125	101896318	0	0.305	*NALCN*	ENSG00000102452	100.0	Variant located inside the gene
INDEL	623	15	39372623	39373245	0	0.172	*RP11-624L4.1*	ENSG00000259345	100.0	Variant located inside the gene
INDEL	953	15	71881673	71882625	0	0.101	*THSD4*	ENSG00000187720	100.0	Variant located inside the gene
CNV	3844	15	76891342	76895185	0	0.253	*SCAPER*	ENSG00000140386	100.0	Variant located inside the gene
CNV	3844	15	76891342	76895185	1	0.012	*SCAPER*	ENSG00000140386	100.0	Variant located inside the gene
CNV	1497	15	91981864	91983360	0	0.283	*RP11-661P17.1*	ENSG00000258551	100.0	Variant located inside the gene
CNV	2252	16	58647399	58649650	0	0.234	*CNOT1*	ENSG00000125107	100.0	Variant located inside the gene
CNV	3386	16	76540062	76543447	0	0.499	*CNTNAP4*	ENSG00000152910	100.0	Variant located inside the gene
CNV	11036	16	78373700	78384735	0	0.242	*WWOX*	ENSG00000186153	100.0	Variant located inside the gene
CNV	11036	16	78373700	78384735	1	0.067	*WWOX*	ENSG00000186153	100.0	Variant located inside the gene
INDEL	360	17	724239	724598	0	0.345	*NXN*	ENSG00000167693	100.0	Variant located inside the gene
CNV	2782	17	35755867	35758648	0	0.145	*ACACA*	ENSG00000132142	100.0	Variant located inside the gene
CNV	2782	17	35755867	35758648	1	0.010	*ACACA*	ENSG00000132142	100.0	Variant located inside the gene

CNV	1677	17	55688120	55689796	0	0.360	*MSI2*	ENSG00000153944	100.0	Variant located inside the gene
INDEL	518	18	24571673	24572190	0	0.168	*AQP4-AS1*	ENSG00000260372	100.0	Variant located inside the gene
INDEL	518	18	24571673	24572190	0	0.168	*CHST9*	ENSG00000154080	100.0	Variant located inside the gene
CNV	3166	18	47695103	47698268	0	0.107	*MYO5B*	ENSG00000167306	100.0	Variant located inside the gene
INDEL	930	18	75267039	75267968	0	0.325	*RP11-176N18.2*	ENSG00000264015	100.0	Variant located inside the gene
INDEL	727	19	2909643	2910369	0	0.794	*ZNF57*	ENSG00000171970	100.0	Variant located inside the gene
CNV	2427	19	12694963	12697389	0	0.483	*ZNF490*	ENSG00000188033	100.0	Variant located inside the gene
CNV	1676	21	19327135	19328810	0	0.105	*CHODL*	ENSG00000154645	100.0	Variant located inside the gene
CNV	1676	21	19327135	19328810	1	0.184	*CHODL*	ENSG00000154645	100.0	Variant located inside the gene
CNV	2812	21	44970373	44973184	0	0.303	*RPL31P1*	ENSG00000214326	13.2	Variant covers the whole gene
CNV	2812	21	44970373	44973184	0	0.303	*HSF2BP*	ENSG00000160207	100.0	Variant located inside the gene
CNV	1664	22	18058001	18059664	0	0.796	*SLC25A18*	ENSG00000182902	100.0	Variant located inside the gene
CNV	2471	22	24365041	24367511	0	0.139	*AP000351.9*	ENSG00000184490	30.8	Variant partially overlaps with the gene
INDEL	529	22	35645524	35646052	0	0.428	*RNU7-167P*	ENSG00000238584	11.6	Variant covers the whole gene

CHR: chromosome; CN: copy number state; freq: frequency

Supplementary Table 7. Multivariable Cox PH (proportional hazards) regression analysis results assuming that all variables satisfy the PH assumption

Gene	INDEL/CNV	p-value	HR	95% CI (lower)	95% CI (higher)
PADI4	CHR_1_17676291_17677196 (0 copy vs 2 copy)	0.786	1.067	0.669	1.701
DAB1	CHR_1_58744143_58744663 (0 copy vs 2 copy)	0.92	0.983	0.7	1.379
RPL31P12	CHR_1_72766413_72811692 (0 copy vs 2 copy)	0.846	0.97	0.716	1.315
GBP3	CHR_1_89476427_89478432 (0 copy vs 2 copy)	0.89	1.033	0.652	1.638
TGFBR3	CHR_1_92232111_92233227 (0 copy vs 2 copy)	**0.033**	0.5	0.264	0.946
LCE3C, LCE3B	CHR_1_152556085_152586939 (0 copy vs 2 or 3 copy)	0.942	0.988	0.719	1.359
NME7	CHR_1_169207360_169241309 (0 copy vs 1 or 2 copy)	0.975	0.992	0.615	1.602
TDRD5	CHR_1_179607382_179607936 (0 copy vs 2 copy)	0.701	1.087	0.712	1.659
C4BPA	CHR_1_207292578_207293178 (0 copy vs 2 copy)	0.647	1.069	0.802	1.425
LTBP1	CHR_2_33224605_33227187 (0 copy vs 2 copy)	0.737	1.083	0.681	1.723
AC073218.1	CHR_2_34698447_34736476 (0 copy vs 1 or 2 copy)	0.172	1.306	0.89	1.916
C2orf73	CHR_2_54565729_54567441 (0 copy vs 2 copy)	0.637	0.929	0.685	1.26
C2orf73	CHR_2_54565729_54567590 (0 copy vs 1 or 2 copy)	0.57	0.908	0.651	1.267
SNRNP27, MXD1	CHR_2_70125092_70125504 (0 copy vs 2 copy)	0.313	0.805	0.529	1.226
REV1	CHR_2_100103752_100105013 (0 copy vs 2 copy)	0.158	1.284	0.908	1.817
TANC1	CHR_2_159959587_159961014 (0 copy vs 2 copy)	0.264	0.832	0.602	1.149
TANC1	CHR_2_159959587_159961451 (0 copy vs 2 copy)	0.747	0.936	0.628	1.396
PPP1R1C	CHR_2_182856938_182857477 (0 copy vs 2 copy)	0.446	1.187	0.764	1.843
AC072062.1	CHR_2_215728845_215730688 (0 copy vs 2 copy)	0.965	0.991	0.644	1.523
NIFKP7, OSBPL10	CHR_3_32102055_32106725 (0 copy vs 1 or 2 copy)	0.469	1.138	0.802	1.616
SCAP	CHR_3_47490712_47493338 (0 copy vs 2 copy)	0.211	0.769	0.509	1.161
CMSS1, FILIP1L	CHR_3_99628822_99629567 (0 copy vs 1 or 2 copy)	**0.019**	1.661	1.087	2.536
LINC00883, RP11-446H18.5	CHR_3_107038162_107040253 (0 copy vs 2 copy)	0.848	0.956	0.605	1.512
SLC12A8	CHR_3_124936371_124936911 (0 copy vs 1 or 2 copy)	0.853	1.036	0.713	1.504

CPNE4	CHR_3_131708352_131713017 (0 copy vs 1 or 2 copy)	0.173	1.325	0.884	1.985
PCCB	CHR_3_136021052_136026101 (0 copy vs 1 or 2 copy)	0.659	1.102	0.716	1.696
IQCJ-SCHIP1	CHR_3_159257057_159257610 (0 copy vs 2 copy)	0.514	1.161	0.742	1.815
RP11-10O22.1	CHR_3_162765807_162769007 (0 copy vs 2 copy)	0.792	0.962	0.722	1.282
LEPREL1	CHR_3_189737354_189740440 (0 copy vs 1 or 2 copy)	0.096	1.441	0.938	2.215
RP11-143P4.2	CHR_3_192875738_192885153 (0 copy vs 2 or 4 copy)	0.059	1.321	0.99	1.764
CCSER1	CHR_4_91933043_91935779 (0 copy vs 2 copy)	0.719	0.922	0.593	1.435
NDST4	CHR_4_115928747_115929279 (0 copy vs 2 copy)	0.111	0.632	0.359	1.112
LINC00616	CHR_4_138966505_138967151 (0 copy vs 2 copy)	0.82	0.95	0.612	1.475
SMAD1	CHR_4_146438871_146439991 (0 copy vs 1 or 2 copy)	0.94	1.015	0.682	1.511
TMEM192	CHR_4_166003471_166004657 (0 copy vs 2 copy)	0.349	1.183	0.832	1.683
GALNTL6	CHR_4_172989075_172992876 (0 copy vs 2 copy)	0.842	1.035	0.737	1.453
LINC00290	CHR_4_182056607_182057107 (0 copy vs 2 copy)	0.926	1.015	0.744	1.384
PDLIM3	CHR_4_186441932_186444023 (0 copy vs 2 copy)	0.125	0.799	0.599	1.065
PDLIM3	CHR_4_186441932_186444110 (0 copy vs 2 copy)	0.186	1.265	0.893	1.794
FAM149A	CHR_4_187093557_187098071 (0 copy vs 2 copy)	0.831	1.035	0.756	1.415
CTD-3080P12.3	CHR_5_1178511_1180425 (0 copy vs 2 copy)	0.177	1.269	0.898	1.793
GUSBP1	CHR_5_21450792_21452439 (0 copy vs 2 copy)	0.832	0.956	0.634	1.442
CTC-254B4.1	CHR_5_106324802_106326299 (0 copy vs 2 copy)	0.943	1.014	0.702	1.464
SPINK14	CHR_5_147553186_147554186 (0 copy vs 2 copy)	0.764	0.954	0.703	1.296
EXOC2	CHR_6_666535_667756 (0 copy vs 2 copy)	0.927	0.986	0.734	1.326
snoU13, RNF144B	CHR_6_18402172_18402672 (0 copy vs 2 copy)	0.366	1.143	0.855	1.529
PKHD1	CHR_6_51736175_51736742 (0 copy vs 1 or 2 copy)	0.211	1.226	0.891	1.689
MLIP-AS1, MLIP	CHR_6_53929777_53933874 (0 copy vs 2 copy)	0.235	0.839	0.628	1.121
EYS	CHR_6_65347533_65349159 (0 copy vs 2 copy)	0.466	0.89	0.65	1.219
GABRR1	CHR_6_89921782_89922171 (0 copy vs 2 copy)	0.284	1.178	0.873	1.591
RP11-517H2.6	CHR_6_167488211_167489090 (0 copy vs 2 copy)	0.516	1.132	0.778	1.648
STEAP2-AS1, STEAP2	CHR_7_89810608_89811996 (0 copy vs 2 copy)	0.638	0.891	0.551	1.441

STEAP2-AS1, STEAP2	CHR_7_89810608_89812114 (0 copy vs 2 copy)	**0.031**	0.568	0.34	0.95
AC000370.2	CHR_7_126048572_126051369 (0 copy vs 2 copy)	0.444	0.826	0.506	1.348
AC000370.2	CHR_7_126048572_126051476 (0 copy vs 2 copy)	0.739	0.937	0.639	1.374
CNTNAP2	CHR_7_148074379_148076266 (0 copy vs 2 copy)	0.395	0.876	0.647	1.188
ERICH1	CHR_8_594761_599201 (0 copy vs 2 copy)	0.874	0.976	0.72	1.322
CSMD1	CHR_8_4122961_4124156 (0 copy vs 2 copy)	0.904	0.982	0.734	1.314
C8orf12	CHR_8_11245641_11247049 (0 copy vs 2 copy)	0.915	0.98	0.675	1.423
DOCK5	CHR_8_25066884_25070636 (0 copy vs 2 copy)	0.81	1.036	0.774	1.388
ADAM5, ADAM3A	CHR_8_39233344_39387179 (0 copy vs 1 or 2 copy)	0.36	0.831	0.559	1.235
GDAP1	CHR_8_75364528_75366830 (0 copy vs 2 copy)	0.966	1.007	0.747	1.357
RP11-149P24.1	CHR_8_137160319_137163816 (0 copy vs 2 copy)	0.091	0.742	0.525	1.049
TJP2	CHR_9_71741217_71743100 (0 copy vs 2 copy)	0.059	0.756	0.565	1.011
GABBR2	CHR_9_101309058_101311079 (0 copy vs 2 copy)	0.756	1.057	0.745	1.499
VTI1BP4, WDR34	CHR_9_131412549_131413853 (0 copy vs 2 copy)	0.882	0.977	0.719	1.327
VTI1BP4, WDR34	CHR_9_131412549_131413885 (0 copy vs 2 copy)	0.183	0.822	0.616	1.097
RP11-98L5.4	CHR_9_138479177_138480145 (0 copy vs 2 copy)	0.805	1.038	0.772	1.396
LINC00704	CHR_10_4708627_4710298 (0 copy vs 2 copy)	0.724	1.053	0.789	1.406
PDSS1	CHR_10_27000558_27001814 (0 copy vs 2 copy)	0.645	0.932	0.691	1.257
C10orf11	CHR_10_78255873_78260694 (0 copy vs 2 copy)	0.9	1.02	0.752	1.384
MINPP1	CHR_10_89275888_89276407 (0 copy vs 1 or 2 copy)	0.943	0.989	0.738	1.326
LGI1	CHR_10_95545536_95546273 (0 copy vs 2 copy)	0.879	1.04	0.63	1.715
GUCY2GP	CHR_10_114113589_114116575 (0 copy vs 2 copy)	0.096	0.667	0.415	1.074
PPAPDC1A	CHR_10_122226947_122228534 (0 copy vs 2 copy)	0.428	0.842	0.551	1.288
TRIM5	CHR_11_5760106_5762286 (0 copy vs 2 copy)	0.761	1.066	0.707	1.606
TMEM41B	CHR_11_9324025_9324496 (0 copy vs 2 copy)	0.871	0.972	0.692	1.367
DNAJC24	CHR_11_31394060_31397428 (0 copy vs 1 or 2 copy)	0.935	0.983	0.644	1.498
RP11-430H10.4	CHR_11_45430401_45431405 (0 copy vs 2 copy)	0.581	0.914	0.663	1.259
PC	CHR_11_66712229_66713105 (0 copy vs 2 copy)	0.251	1.22	0.869	1.711

ETV6	CHR_12_12026506_12026937 (0 copy vs 2 copy)	0.679	0.903	0.555	1.468
SLC15A5	CHR_12_16420184_16420943 (0 copy vs 2 copy)	0.777	1.056	0.725	1.538
RP11-352M15.1	CHR_12_45903118_45909531 (0 copy vs 1 or 2 copy)	0.668	0.931	0.673	1.289
LHFP	CHR_13_39934551_39935151 (0 copy vs 2 copy)	0.366	0.871	0.645	1.176
DLEU1	CHR_13_51069352_51072600 (0 copy vs 2 copy)	0.827	1.033	0.775	1.376
NALCN	CHR_13_101894125_101896318 (0 copy vs 2 copy)	0.077	0.741	0.532	1.033
RP11-624L4.1	CHR_15_39372623_39373245 (0 copy vs 2 copy)	0.156	1.3	0.905	1.869
THSD4	CHR_15_71881673_71882625 (0 copy vs 2 copy)	0.772	0.936	0.597	1.466
SCAPER	CHR_15_76891342_76895185 (0 copy vs 1 or 2 copy)	0.759	0.949	0.681	1.323
RP11-661P17.1	CHR_15_91981864_91983360 (0 copy vs 2 copy)	0.601	1.086	0.797	1.48
CNOT1	CHR_16_58647399_58649650 (0 copy vs 2 copy)	0.409	1.148	0.827	1.594
CNTNAP4	CHR_16_76540062_76543447 (0 copy vs 2 copy)	0.875	0.977	0.733	1.302
WWOX	CHR_16_78373700_78384735 (0 copy vs 1 or 2 copy)	0.777	0.953	0.681	1.333
NXN	CHR_17_724239_724598 (0 copy vs 2 copy)	0.937	0.988	0.732	1.334
ACACA	CHR_17_35755867_35758648 (0 copy vs 1 or 2 copy)	0.16	0.74	0.487	1.126
MSI2	CHR_17_55688120_55689796 (0 copy vs 2 copy)	0.857	0.973	0.72	1.315
AQP4-AS1, CHST9	CHR_18_24571673_24572190 (0 copy vs 2 copy)	0.74	0.935	0.629	1.391
MYO5B	CHR_18_47695103_47698268 (0 copy vs 2 copy)	0.2	1.323	0.862	2.031
RP11-176N18.2	CHR_18_75267039_75267968 (0 copy vs 2 copy)	0.517	0.902	0.661	1.232
ZNF57	CHR_19_2909643_2910369 (0 copy vs 2 copy)	0.503	1.135	0.783	1.645
ZNF490	CHR_19_12694963_12697389 (0 copy vs 2 copy)	0.82	1.034	0.776	1.377
CHODL	CHR_21_19327135_19328810 (0 copy vs 1 or 2 copy)	0.621	1.125	0.705	1.796
RPL31P1, HSF2BP	CHR_21_44970373_44973184 (0 copy vs 2 copy)	0.824	0.965	0.706	1.32
SLC25A18	CHR_22_18058001_18059664 (0 copy vs 2 copy)	0.876	0.972	0.682	1.387
AP000351.9	CHR_22_24365041_24367511 (0 copy vs 2 copy)	0.338	0.81	0.526	1.247
RNU7-167P	CHR_22_35645524_35646052 (0 copy vs 2 copy)	0.085	0.774	0.577	1.036

CHR: Chromosome; CI: confidence interval; HR: hazards ratio; PH: proportional hazards. Models are adjusted for stage, location, and MSI status. P-values less than 0.05 are shown in bold fonts.

Supplementary Table 8. Results of the age-stratified Multivariable Cox regression analysis for the INDELs/CNVs

Gene	INDEL/CNV	p-value	HR	95% CI (lower)	95% CI (higher)	*p-value for PH
PADI4	CHR_1_17676291_17677196 (0 CN vs 2 CN)	0.7079	1.0935	0.685	1.7455	0.8309
DAB1	CHR_1_58744143_58744663 (0 CN vs 2 CN)	0.9091	0.9804	0.698	1.3771	0.5063
RPL31P12	CHR_1_72766413_72811692 (0 CN vs 2 CN)	0.9304	0.9865	0.7268	1.3389	0.7636
GBP3	CHR_1_89476427_89478432 (0 CN vs 2 CN)	0.9685	1.0094	0.6351	1.6042	0.925
TGFBR3	CHR_1_92232111_92233227 (0 CN vs 2 CN)	**0.0454**	0.5211	0.2752	0.9867	0.2979
LCE3C, LCE3B	CHR_1_152556085_152586939 (0 CN vs 2 or 3 CN)	0.8769	1.0257	0.7438	1.4144	0.2937
NME7	CHR_1_169207360_169241309 (0 CN vs 1 or 2 CN)	0.96	1.0123	0.627	1.6344	**0.0413**
TDRD5	CHR_1_179607382_179607936 (0 CN vs 2 CN)	0.7219	1.0799	0.7072	1.6491	0.0553
C4BPA	CHR_1_207292578_207293178 (0 CN vs 2 CN)	0.6666	1.0652	0.7992	1.4197	0.0741
LTBP1	CHR_2_33224605_33227187 (0 CN vs 2 CN)	0.7589	1.0756	0.6755	1.7125	0.7016
AC073218.1	CHR_2_34698447_34736476 (0 CN vs 1 or 2 CN)	0.1774	1.302	0.8873	1.9105	0.4287
C2orf73	CHR_2_54565729_54567441 (0 CN vs 2 CN)	0.658	0.9334	0.6881	1.2663	0.3946
C2orf73	CHR_2_54565729_54567590 (0 CN vs 1 or 2 CN)	0.6066	0.9161	0.6562	1.2788	0.8349
SNRNP27, MXD1	CHR_2_70125092_70125504 (0 CN vs 2 CN)	0.3241	0.8092	0.5313	1.2325	0.7296
REV1	CHR_2_100103752_100105013 (0 CN vs 2 CN)	0.2135	1.2473	0.8806	1.7666	0.7126
TANC1	CHR_2_159959587_159961014 (0 CN vs 2 CN)	0.2936	0.8409	0.6086	1.1619	0.4102
TANC1	CHR_2_159959587_159961451 (0 CN vs 2 CN)	0.6495	0.9113	0.6106	1.3602	0.2303
PPP1R1C	CHR_2_182856938_182857477 (0 CN vs 2 CN)	0.4582	1.1819	0.7601	1.8377	0.4066
AC072062.1	CHR_2_215728845_215730688 (0 CN vs 2 CN)	0.9146	0.9767	0.6346	1.5031	0.6646
NIFKP7, OSBPL10	CHR_3_32102055_32106725 (0 CN vs 1 or 2 CN)	0.5554	1.1116	0.7821	1.5798	0.6074
SCAP	CHR_3_47490712_47493338 (0 CN vs 2 CN)	0.2445	0.7824	0.5175	1.1828	0.0841
CMSS1, FILIP1L	CHR_3_99628822_99629567 (0 CN vs 1 or 2 CN)	**0.015**	1.6936	1.1076	2.5896	0.3444
LINC00883, RP11-446H18.5	CHR_3_107038162_107040253 (0 CN vs 2 CN)	0.8988	0.9707	0.6133	1.5362	0.8449
SLC12A8	CHR_3_124936371_124936911 (0 CN vs 1 or 2 CN)	0.8107	1.0467	0.7204	1.5208	0.573
CPNE4	CHR_3_131708352_131713017 (0 CN vs 1 or 2 CN)	0.1825	1.3172	0.8785	1.9751	0.7708

PCCB	CHR_3_136021052_136026101 (0 CN vs 1 or 2 CN)	0.7009	1.0883	0.7067	1.6761	0.1133
IQCJ-SCHIP1	CHR_3_159257057_159257610 (0 CN vs 2 CN)	0.4461	1.191	0.7597	1.8673	0.6704
RP11-10O22.1	CHR_3_162765807_162769007 (0 CN vs 2 CN)	0.6661	0.9383	0.7025	1.2531	0.287
LEPREL1	CHR_3_189737354_189740440 (0 CN vs 1 or 2 CN)	0.1296	1.3983	0.9064	2.1569	0.8318
RP11-143P4.2	CHR_3_192875738_192885153 (0 CN vs 2 or 4 CN)	**0.0394**	1.3586	1.0149	1.8186	0.9002
CCSER1	CHR_4_91933043_91935779 (0 CN vs 2 CN)	0.7505	0.9307	0.5981	1.4485	0.125
NDST4	CHR_4_115928747_115929279 (0 CN vs 2 CN)	0.1438	0.6551	0.3715	1.1551	0.8987
LINC00616	CHR_4_138966505_138967151 (0 CN vs 2 CN)	0.9368	0.9823	0.6315	1.5279	0.9605
SMAD1	CHR_4_146438871_146439991 (0 CN vs 1 or 2 CN)	0.8007	1.0528	0.7062	1.5695	0.469
TMEM192	CHR_4_166003471_166004657 (0 CN vs 2 CN)	0.3689	1.1755	0.8261	1.6727	0.548
GALNTL6	CHR_4_172989075_172992876 (0 CN vs 2 CN)	0.823	1.0397	0.7393	1.462	0.733
LINC00290	CHR_4_182056607_182057107 (0 CN vs 2 CN)	0.8276	1.0352	0.7583	1.4131	0.7641
PDLIM3	CHR_4_186441932_186444023 (0 CN vs 2 CN)	0.1215	0.7969	0.5978	1.0622	0.1884
PDLIM3	CHR_4_186441932_186444110 (0 CN vs 2 CN)	0.2439	1.2318	0.8674	1.7493	**0.0119**
FAM149A	CHR_4_187093557_187098071 (0 CN vs 2 CN)	0.7014	1.0633	0.7771	1.4548	0.2327
CTD-3080P12.3	CHR_5_1178511_1180425 (0 CN vs 2 CN)	0.2506	1.2269	0.8655	1.7391	0.9494
GUSBP1	CHR_5_21450792_21452439 (0 CN vs 2 CN)	0.7988	0.9479	0.6284	1.4299	**0.0218**
CTC-254B4.1	CHR_5_106324802_106326299 (0 CN vs 2 CN)	0.9655	0.9919	0.6863	1.4336	0.2209
SPINK14	CHR_5_147553186_147554186 (0 CN vs 2 CN)	0.6513	0.9316	0.6851	1.2668	0.5383
EXOC2	CHR_6_666535_667756 (0 CN vs 2 CN)	0.7092	0.945	0.7019	1.2722	0.3524
snoU13, RNF144B	CHR_6_18402172_18402672 (0 CN vs 2 CN)	0.427	1.1253	0.841	1.5057	0.966
PKHD1	CHR_6_51736175_51736742 (0 CN vs 1 or 2 CN)	0.2369	1.2143	0.8803	1.675	0.3962
MLIP-AS1, MLIP	CHR_6_53929777_53933874 (0 CN vs 2 CN)	0.2985	0.8569	0.6405	1.1465	0.9597
EYS	CHR_6_65347533_65349159 (0 CN vs 2 CN)	0.5346	0.905	0.6606	1.24	0.6478
GABRR1	CHR_6_89921782_89922171 (0 CN vs 2 CN)	0.2558	1.1904	0.8813	1.6079	0.9468
RP11-517H2.6	CHR_6_167488211_167489090 (0 CN vs 2 CN)	0.4816	1.1451	0.7851	1.6701	0.8693
STEAP2-AS1, STEAP2	CHR_7_89810608_89811996 (0 CN vs 2 CN)	0.6569	0.8966	0.554	1.4512	0.5303
STEAP2-AS1, STEAP2	CHR_7_89810608_89812114 (0 CN vs 2 CN)	**0.0372**	0.5776	0.3447	0.968	0.4002

AC000370.2	CHR_7_126048572_126051369 (0 CN vs 2 CN)	0.3971	0.8089	0.4951	1.3216	0.1919
AC000370.2	CHR_7_126048572_126051476 (0 CN vs 2 CN)	0.8387	0.961	0.6549	1.41	0.2056
CNTNAP2	CHR_7_148074379_148076266 (0 CN vs 2 CN)	0.4389	0.8864	0.6531	1.203	0.7985
ERICH1	CHR_8_594761_599201 (0 CN vs 2 CN)	0.6277	0.9268	0.6818	1.26	0.6751
CSMD1	CHR_8_4122961_4124156 (0 CN vs 2 CN)	0.8471	0.9717	0.7254	1.3015	0.6223
C8orf12	CHR_8_11245641_11247049 (0 CN vs 2 CN)	0.998	1.0005	0.6885	1.4538	0.9594
DOCK5	CHR_8_25066884_25070636 (0 CN vs 2 CN)	0.7665	1.0454	0.7799	1.4011	0.9903
ADAM5, ADAM3A	CHR_8_39233344_39387179 (0 CN vs 1 or 2 CN)	0.3411	0.8248	0.5547	1.2263	0.8655
GDAP1	CHR_8_75364528_75366830 (0 CN vs 2 CN)	0.8082	1.038	0.7683	1.4022	0.1561
RP11-149P24.1	CHR_8_137160319_137163816 (0 CN vs 2 CN)	0.0786	0.7328	0.5182	1.0362	0.1977
TJP2	CHR_9_71741217_71743100 (0 CN vs 2 CN)	0.0581	0.7543	0.5636	1.0097	0.0958
GABBR2	CHR_9_101309058_101311079 (0 CN vs 2 CN)	0.652	1.0839	0.7637	1.5386	0.5429
VTI1BP4, WDR34	CHR_9_131412549_131413853 (0 CN vs 2 CN)	0.9026	0.9811	0.7223	1.3325	0.7366
VTI1BP4, WDR34	CHR_9_131412549_131413885 (0 CN vs 2 CN)	0.1921	0.8249	0.6178	1.1015	0.866
RP11-98L5.4	CHR_9_138479177_138480145 (0 CN vs 2 CN)	0.8617	1.0268	0.7627	1.3823	0.9731
LINC00704	CHR_10_4708627_4710298 (0 CN vs 2 CN)	0.7502	1.0482	0.7847	1.4001	0.7738
PDSS1	CHR_10_27000558_27001814 (0 CN vs 2 CN)	0.5513	0.9123	0.6745	1.2339	0.4973
C10orf11	CHR_10_78255873_78260694 (0 CN vs 2 CN)	0.757	1.0496	0.7726	1.4258	0.5459
MINPP1	CHR_10_89275888_89276407 (0 CN vs 1 or 2 CN)	0.896	1.0198	0.7604	1.3675	0.7792
LGI1	CHR_10_95545536_95546273 (0 CN vs 2 CN)	0.9451	1.0177	0.6168	1.6794	0.9142
GUCY2GP	CHR_10_114113589_114116575 (0 CN vs 2 CN)	0.0874	0.6595	0.4091	1.063	0.1796
PPAPDC1A	CHR_10_122226947_122228534 (0 CN vs 2 CN)	0.3876	0.8295	0.5429	1.2676	0.2956
TRIM5	CHR_11_5760106_5762286 (0 CN vs 2 CN)	0.9335	1.0177	0.6746	1.5351	0.8085
TMEM41B	CHR_11_9324025_9324496 (0 CN vs 2 CN)	0.9633	0.992	0.7056	1.3947	0.8368
DNAJC24	CHR_11_31394060_31397428 (0 CN vs 1 or 2 CN)	0.9041	0.9744	0.6385	1.4869	0.3131
RP11-430H10.4	CHR_11_45430401_45431405 (0 CN vs 2 CN)	0.5584	0.9086	0.6592	1.2525	0.0612
PC	CHR_11_66712229_66713105 (0 CN vs 2 CN)	0.2875	1.2015	0.8566	1.6852	0.7789
ETV6	CHR_12_12026506_12026937 (0 CN vs 2 CN)	0.9327	0.9792	0.6007	1.5962	0.5481

SLC15A5	CHR_12_16420184_16420943 (0 CN vs 2 CN)	0.7365	1.0668	0.7319	1.5551	0.4241
RP11-352M15.1	CHR_12_45903118_45909531 (0 CN vs 1 or 2 CN)	0.6838	0.9347	0.6754	1.2936	0.7158
LHFP	CHR_13_39934551_39935151 (0 CN vs 2 CN)	0.3578	0.8681	0.6422	1.1735	0.3508
DLEU1	CHR_13_51069352_51072600 (0 CN vs 2 CN)	0.8121	1.0354	0.777	1.3798	0.4998
NALCN	CHR_13_101894125_101896318 (0 CN vs 2 CN)	0.0798	0.7427	0.5324	1.0359	0.2969
RP11-624L4.1	CHR_15_39372623_39373245 (0 CN vs 2 CN)	0.2274	1.2515	0.8694	1.8015	0.7193
THSD4	CHR_15_71881673_71882625 (0 CN vs 2 CN)	0.7324	0.9243	0.5885	1.4516	0.0673
SCAPER	CHR_15_76891342_76895185 (0 CN vs 1 or 2 CN)	0.7548	0.9483	0.6795	1.3235	0.956
RP11-661P17.1	CHR_15_91981864_91983360 (0 CN vs 2 CN)	0.5408	1.1017	0.8077	1.5026	0.2388
CNOT1	CHR_16_58647399_58649650 (0 CN vs 2 CN)	0.3685	1.163	0.8369	1.616	0.1922
CNTNAP4	CHR_16_76540062_76543447 (0 CN vs 2 CN)	0.7307	0.9505	0.7118	1.2692	0.4307
WWOX	CHR_16_78373700_78384735 (0 CN vs 1 or 2 CN)	0.9201	0.9829	0.701	1.378	0.8013
NXN	CHR_17_724239_724598 (0 CN vs 2 CN)	0.8559	0.9726	0.7207	1.3126	0.8732
ACACA	CHR_17_35755867_35758648 (0 CN vs 1 or 2 CN)	0.1951	0.7572	0.4971	1.1534	0.1903
MSI2	CHR_17_55688120_55689796 (0 CN vs 2 CN)	0.8083	0.9634	0.7128	1.302	0.7803
AQP4-AS1, CHST9	CHR_18_24571673_24572190 (0 CN vs 2 CN)	0.6977	0.9242	0.621	1.3756	0.1238
MYO5B	CHR_18_47695103_47698268 (0 CN vs 2 CN)	0.2691	1.2759	0.8282	1.9655	0.3915
RP11-176N18.2	CHR_18_75267039_75267968 (0 CN vs 2 CN)	0.4543	0.8876	0.6494	1.2131	0.8128
ZNF57	CHR_19_2909643_2910369 (0 CN vs 2 CN)	0.5232	1.1288	0.7782	1.6375	0.4248
ZNF490	CHR_19_12694963_12697389 (0 CN vs 2 CN)	0.8367	1.0307	0.7732	1.3739	0.2057
CHODL	CHR_21_19327135_19328810 (0 CN vs 1 or 2 CN)	0.5629	1.1483	0.7187	1.8347	0.5398
RPL31P1, HSF2BP	CHR_21_44970373_44973184 (0 CN vs 2 CN)	0.8447	0.9691	0.7082	1.3263	0.6559
SLC25A18	CHR_22_18058001_18059664 (0 CN vs 2 CN)	0.8535	0.967	0.6776	1.3801	0.4156
AP000351.9	CHR_22_24365041_24367511 (0 CN vs 2 CN)	0.2914	0.7927	0.5148	1.2205	0.1053
RNU7-167P	CHR_22_35645524_35646052 (0 CN vs 2 CN)	0.0626	0.7569	0.5646	1.0148	0.2091

CHR: Chromosome; CI: confidence interval; HR: hazards ratio; PH: proportional hazards. Models are stratified for age and adjusted for stage, location, and MSI status. *p-value by the Score test [38]. P-values less than 0.05 are shown in bold fonts.

Variants that deviated from the PH assumption are shown with green highlights; these results are only shown for the purpose of the comparison of the results obtained with or without the time-varying coefficients, and the results considering the time-varying coefficients shown in the manuscript should be considered more accurate.

Supplementary Table 9. INDELs/CNVs that are unique to FCCX cases.

CHR	START position	END position	*CHR band	CN	Length	ENSEMBL ID	Gene
1	104100911	104136650	1p21.1	3	35739	ENSG00000236085, ENSG00000240038	*ACTG1P4, AMY2B*
1	108858527	108888654	1p13.3	3	30127	ENSG00000241361	*SLC25A24P1*
2	1527274	1537864	2p25.3	3	10590	ENSG00000115705	*TPO*
2	24602959	24611360	2p23.3	0	8401	-	-
2	34698447	34729435	2p22.3	0	30988	ENSG00000226785	*AC073218.1*
2	90125210	90283546	2p11.2	3	158336	ENSG00000178894, ENSG00000254292, ENSG00000253906, ENSG00000241244, ENSG00000224041, ENSG00000211630, ENSG00000240834, ENSG00000211632, ENSG00000211633, ENSG00000242580, ENSG00000239819, ENSG00000235896	*AC073416.1, IGKV2D-14, IGKV2D-10, IGKV1D-16, IGKV3D-15, IGKV1D-13, IGKV1D-12, IGKV3D-11, IGKV1D-42, IGKV1D-43, IGKV1D-8, IGKV3D-7*
2	105662069	105665808	2q12.1	0	3739	ENSG00000135972	*MRPS9*
3	131708352	131712742	3q22.1	1	4390	ENSG00000196353	*CPNE4*
3	162509817	162656280	3q22.1	4	146463	-	-
3	195453991	195472740	3q26.1	3	18749	ENSG00000242086, ENSG00000176945	*LINC00969, MUC20*
4	10272429	10274279	4p16.1	1	1850	-	-
6	29862114	29890271	6p22.1	1	28157	ENSG00000231130, ENSG00000233677, ENSG00000235963	*HLA-T,DDX39BP1, MCCD1P1*
6	29879173	29905193	6p22.1	3	26020	ENSG00000227262, ENSG00000230795, ENSG00000228078	*HCG4B, HLA-K, HLA-U*
7	87669005	87672670	7q21.12	1	3665	ENSG00000008277	*ADAM22*

8	11924124	12010237	8p23.1	3	86113	ENSG00000252029, ENSG00000255544, ENSG00000255052, ENSG00000254923, ENSG00000226430, ENSG00000254866, ENSG00000233050, ENSG00000215343, ENSG00000223443	*RNA5SP253, DEFB108P3, FAM66D, RP11-1236K1.8, USP17L7, DEFB109P3, DEFB130, ZNF705D, USP17L2*
8	12232330	12251955	8p23.1	3	19625	ENSG00000227888, ENSG00000254423, ENSG00000255556, ENSG00000242296	*FAM66A, RP11-351I21.7, RP11-351I21.6, DEFB109P1*
9	75800805	75808530	9q21.13	1	7725	-	-
10	89007613	89108950	10q23.2	1	101337	ENSG00000223482, ENSG00000224914	*NUTM2A-AS1, LINC00863*
11	93683453	93688134	11q21	1	4681	-	-
12	7997547	8116068	12p13.31	3	118521	ENSG00000173262, ENSG00000059804, ENSG00000222978, ENSG00000201663, ENSG00000241828, ENSG00000255885, ENSG00000176654, ENSG00000255356	*SLC2A14, SLC2A3, Y_RNA, Y_RNA, RP11-277J24.1, RP11-815D16.1, NANOGP1, RP11-277E18.2*
13	57787187	57788023	13q21.1	0	836	-	-
15	22521113	22560308	15q11.2	1	39195	ENSG00000259098	*RP11-603B24.2*
15	24497572	24694058	15q11.2	3	196486	ENSG00000261621, ENSG00000261598, ENSG00000260760	*RP11-580I1.2, RP11-107D24.2, PWRN3*
15	56790539	56811204	15q21.3	1	20665	-	-
16	33434576	33632860	16p11.2	3	198284	ENSG00000260518, ENSG00000261580, ENSG00000260308, ENSG00000261153,	*BMS1P8, ENPP7P13, RP11-104C4.4, RP11-104C4.2, IGHV3OR16-12, IGHV3OR16-13*

						ENSG00000270467, ENSG00000271178	
16	78876980	78877618	16q23.1	0	638	ENSG00000186153	*WWOX*
17	44249096	44283571	17q21.31	3	34475	ENSG00000120071, ENSG00000214401	*KANSL1, KANSL1-AS1*
18	61840388	61982829	18q22.1	3	142441	ENSG00000267134, ENSG00000266952	*RP11-146N18.1, RP11-909B2.1*

CHR: chromosome; CN: copy number state. *Based on UCSC genome browser (hg19) (1)

Supplementary Table 10. INDELs/CNVs and genes unique to FCCX cases prior to filtering out the variants based on previous studies

CHR	START	END	*CHR band	CN	Length	ENSEMBL ID	Gene
1	2576908	2785671	1p36.32	0	208763	ENSG00000215912, ENSG00000233234, ENSG00000231630	TTC34, RP11-740P5.2, RP11-740P5.3
1	63041800	63114560	1p31.3	3	72760	ENSG00000213703, ENSG00000116641, ENSG00000132855, ENSG00000269624	RP5-849H19.2, DOCK7, ANGPTL3, AL138847.1
1	92234553	92642484	1p22.1	3	407931	ENSG00000239794, ENSG00000266532, ENSG00000233228, ENSG00000230667, ENSG00000224678, ENSG00000233401, ENSG00000069702, ENSG00000137948, ENSG00000172031, ENSG00000189195, ENSG00000069712	RN7SL653P, RN7SL235P, LPCAT2BP, SETSIP, GAPDHP46, PRKAR1AP, TGFBR3, BRDT, EPHX4, BTBD8, KIAA1107
2	44742517	44759941	2p21	1	17424	ENSG00000143919	CAMKMT
2	182030229	182169888	2q31.3	1	139659	ENSG00000234663	AC104820.2
2	228575393	228607356	2q36.3	3	31963	ENSG00000135917	SLC19A3
3	5419668	5470493	3p26.1	1	50825	-	-
3	82863216	82873921	3p12.2	1	10705	-	-
4	1077575	1083914	4p16.3	3	6339	ENSG00000178222	RNF212
4	9966771	9974186	4p16.1	3	7415	ENSG00000109667	SLC2A9
4	80357377	80362679	4q21.21	1	5302	-	-

4	86789668	86871282	4q21.23	3	81614	ENSG00000265774, ENSG00000138639	AC098870.1, ARHGAP24
4	140115744	140179888	4q31.1	3	64144	ENSG00000207384, ENSG00000252362, ENSG00000206722	Y_RNA, RNU6-506P, RNU6-1074P
4	177006484	177101060	4q34.2	3	94576	ENSG00000201516, ENSG00000150627	SNORA51, WDR17
5	78106062	78111731	5q14.1	1	5669	ENSG00000113273	ARSB
5	128926595	129030478	5q23.3	1	103883	ENSG00000251680, ENSG00000145808	CTC-575N7.1, ADAMTS19
6	74828682	74833405	6q13	1	4723	ENSG00000223786	RP11-554D15.1
6	131074746	131121992	6q23.1	1	47246	-	-
6	142997132	143118715	6q24.2	3	121583	ENSG00000010818, ENSG00000233138, ENSG00000237851	HIVEP2, RP1-67K17.3, RP1-67K17.4
7	5883055	5919298	7p22.1	1	36243	ENSG00000235944, ENSG00000265040	ZNF815P, RN7SL556P
7	72169613	72255343	7q11.22-q11.23	3	85730	ENSG00000270694, ENSG00000270555, ENSG00000254184	RP11-535E8.2, RP11-1394O16.1, TYW1B
11	11875747	11893217	11p15.3	1	17470	ENSG00000255492, ENSG00000170242	CTD-2381F24.1, USP47
11	127748894	127761666	11q24.2	1	12772	-	-
13	108041621	108055161	13q33.3	1	13540	ENSG00000204442	FAM155A
14	90390739	90470981	14q32.11	1	80242	ENSG00000259053, ENSG00000140025, ENSG00000042088	RP11-33N16.3, EFCAB11, TDP1
15	78865893	78874073	15q25.1	3	8180	ENSG00000169684	CHRNA5

16	12826445	12880446	16p13.12	1	54001	ENSG00000260378, ENSG00000261158, ENSG00000103381	CTD-2583P5.1, CTD-2583P5.3, CPPED1
16	15066052	15221957	16p13.11	3	155905	ENSG00000261819, ENSG00000238728, ENSG00000260872, ENSG00000260735, ENSG00000188599, ENSG00000270580, ENSG00000250251, ENSG00000179889, ENSG00000157045, ENSG00000085721	RP11-680G24.4, MIR1972-1, RP11-680G24.5, RP11-72I8.1, NPIPP1, RP11-1186N24.5, PKD1P6, PDXDC1, NTAN1, RRN3
16	16246164	16261251	16p13.11	1	15087	ENSG00000091262	ABCC6
18	51124626	51137738	18q21.2	1	13112	ENSG00000242945	RPL29P32

CHR: chromosome; CN: copy number state. *Based on UCSC genome browser (hg19) [13].

Appendix D: Supporting information for "The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying associations" (Chapter 4)

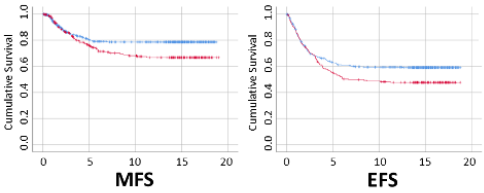
Supplementary Table 11. Pair-wise Pearson correlation coefficient values for the baseline variables.

	Sex	**Histology**	**Location**	**Stage**	**Grade**	**Familial risk**	**MSI status**	**BRAF mutation status**	**Adjuvant chemotherapy**	**Adjuvant radiotherapy**
Sex	1	-0.045	0.111	0.028	0.002	-0.001	-0.095	-0.184	0.065	0.094
Histology		1	-0.069	0.103	0.11	0.030	0.082	0.107	0.038	-0.062
Location			1	-0.082	-0.018	-0.028	-0.175	-0.231	0.235	0.663
Stage				1	0.179	0.038	-0.117	0.037	0.127	0.003
Grade					1	0.012	0.077	0.156	0.021	0.030
Familial risk						1	0.1	0.033	0.060	0.042
MSI status							1	0.368	-0.011	-0.085
BRAF mutation status								1	-0.017	-0.153
Adjuvant chemotherapy									1	0.535
Adjuvant radiotherapy										1

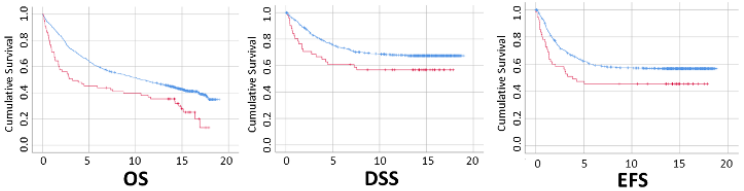
MSI: microsatellite instability.

Supplementary Figures 4-6. Non-crossing curves of Type A variables (**Supplementary Figure 4**) suggest only/mostly protective or detrimental effects but with fluctuating (e.g. increased or decreased) hazard ratios during the follow up while crossing curves of Type B variables (**Supplementary Figure 5**) indicate the changed direction of effects (either from protective effect to detrimental effect, or vice versa). These curve patterns are interesting as some of these variables have their effect directions change over time (e.g. BRAF Val600Glu mutation status in DSS), or have their curves clearly separate only during particular time periods (e.g. adjuvant chemotherapy status in EFS) (**Supplementary Figure 5**). However, in the absence of an assessment for PH assumption by a proper statistical test, interpretation of Kaplan Meier curve patterns may present themselves as a challenge for the researcher. As **Supplementary Figure 5** shows, the crossing nature of the curves may be an initial diagnostics for potential variables, yet for those variables where the curves do not cross (**Supplementary Figure 4**), it is more difficult to make an assessment on whether the variable violates the PH assumption. Thus, as also indicated by others (e.g. Quantin et al. 1999 [39]), in this study a formal assessment of the violation of the PH assumption in Cox models helped identify the variables with time-varying associations.

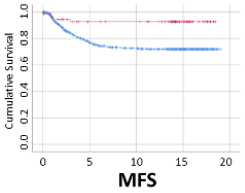
Tumor location
Red – rectum
Blue - colon



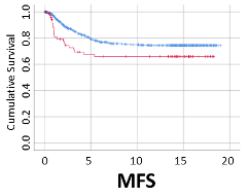
Grade
Red – poorly differentiated
Blue – well/moderately differentiated



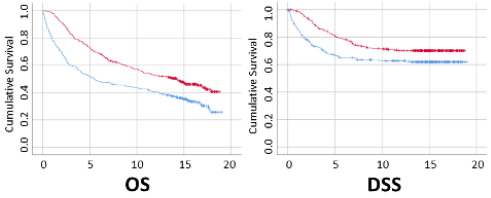
MSI status
Red – MSI-H
Blue – MSI-L/MSS



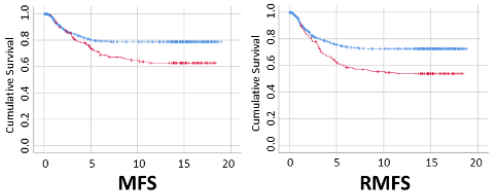
BRAF Val600Glu mutation
Red – yes
Blue – no



Adjuvant chemotherapy treatment
Red – yes
Blue - no

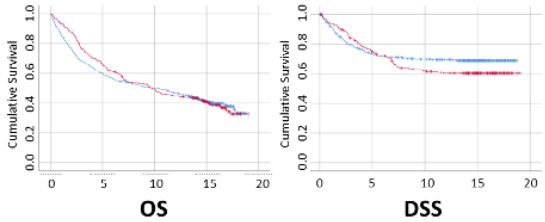


Adjuvant radiotherapy treatment
Red – yes
Blue - no

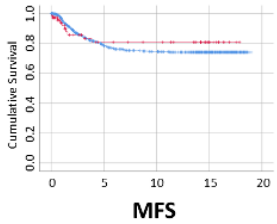


Supplementary Figure 4. Kaplan Meier curves for the variables with a p-value < 0.05 in the univariate Cox analyses and with a p-value < 0.05 in the PH assumption test (**Type A variables**). DSS, disease-specific survival; EFS, event-free survival; MFS, metastasis-free survival; MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; OS, overall survival; RMFS, recurrence/metastasis-free survival. X-axis shows time in years.

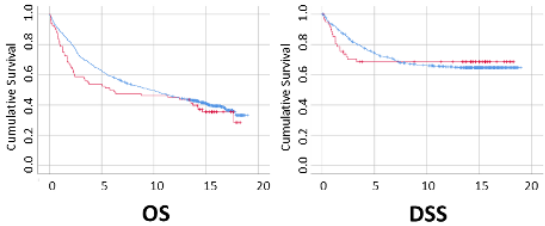
Tumor location
Red – rectum
Blue - colon



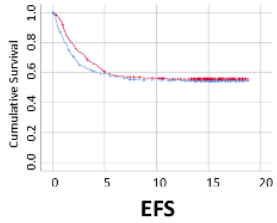
Grade
Red – poorly differentiated
Blue – well/moderately differentiated



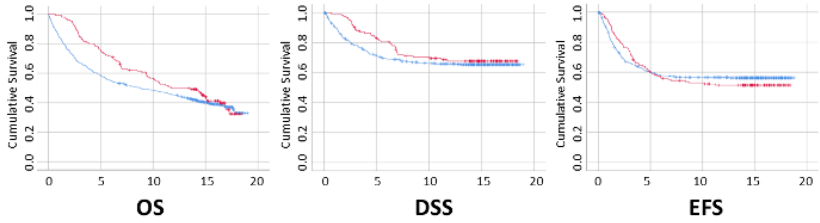
BRAF Val600Glu mutation
Red – yes
Blue – no



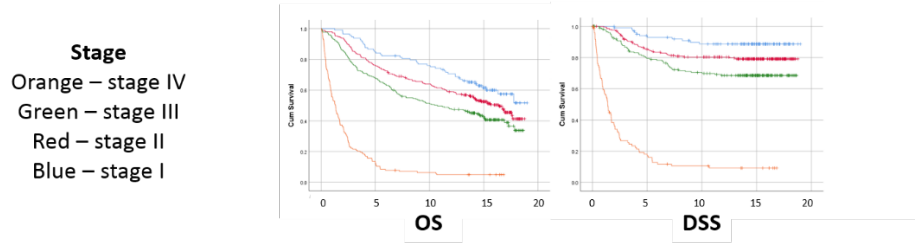
Adjuvant chemotherapy treatment
Red – yes
Blue - no



Adjuvant radiotherapy treatment
Red – yes
Blue - no



Supplementary Figure 5. Kaplan-Meier curves for the variables with a p-value ≥ 0.05 in the univariate Cox analyses and with a p-value < 0.05 in the PH assumption test (**Type B variables**). DSS, disease-free survival; EFS, event-free survival; MFS, metastasis-free survival; OS, overall survival. X-axis shows time in years.



Supplementary Figure 6. Kaplan-Meier curves for disease stage. The stage III and IV subgroups have p-values < 0.05 in the univariate Cox analysis as well as in the PH assumption tests in the OS analysis, and the stage IV subgroup has the p-value < 0.05 in the univariate Cox analysis as well as in the PH assumption test in the DSS analysis (Type A variable). DSS, disease-specific survival; OS, overall survival. X-axis shows time in years. In this variable, not all variable groups (stage II-IV) violated the PH assumption.

Supplementary Table 12. Associations between clinico-demographic/molecular variables and overall survival (OS) in multivariable analysis.

	Cut-off time point T (year)	Time interval	HR	95% CI for HR (lower)	95% CI for HR (upper)	p-value	p-value for PH assumption test
Age at diagnosis	10.5	Before T	1.02	1.01	1.04	**1.35E-04**	0.63
		After T	1.14	1.10	1.19	**1.94E-10**	0.77
Stage (II vs I)			1.88	1.28	2.77	**1.38E-03**	0.21
Stage (III vs I)	1	Before T	38.24	12.31	118.76	**2.94E-10**	0.98
		After T	3.14	1.98	4.98	**1.04E-06**	0.07
Stage (IV vs I)	1	Before T	52.55	20.17	136.95	**5.55E-16**	0.83
		After T	11.83	7.68	18.21	**<2.00E-16**	0.11
Location (rectum vs colon)	2	Before T	0.79	0.50	1.25	0.32	0.99
		After T	1.68	1.29	2.18	**1.23E-04**	0.88
BRAF Val600Glu mutation status (mutant vs wild-type)	2.5	Before T	2.18	1.47	3.23	**9.54E-05**	0.65
		After T	0.70	0.43	1.13	0.15	0.70
Adjuvant chemotherapy treatment (yes vs no)	1	Before T	0.05	0.02	0.13	**1.90E-09**	0.70
		After T	0.56	0.41	0.75	**1.07E-04**	0.18

CI, confidence interval; HR, hazard ratio; PH, proportional hazard. During model construction for OS, the time cut-off points for stage III and adjuvant chemotherapy were estimated at the same time because a satisfying cut-off time point could not be found for stage III when it was analyzed alone (stage III had the smallest p-value of the PH assumption test on OS and adjuvant chemotherapy had the second smallest one. Testing all the possible combinations of the cut-off time points for these two variables identified one year as the proper cut-off point).

Supplementary Table 13. Associations between clinico-demographic/molecular variables and disease-specific survival (DSS) in multivariable analysis.

	Cut-off time point T (year)	Time interval	HR	95% CI for HR (lower)	95% CI for HR (upper)	p-value	p-value for PH assumption test
Stage (II vs I)			2.51	1.25	5.03	**9.46E-03**	0.33
Stage (III vs I)			4.99	2.34	10.62	**3.06E-05**	0.13
Stage (IV vs I)	1	Before T	85.43	25.16	290.03	**9.91E-13**	0.39
		After T	21.69	10.78	43.61	**<2.00E-16**	0.14
Location (rectum vs colon)	6.5	Before T	1.42	0.99	2.03	0.06	0.52
		After T	5.97	2.56	13.93	**3.59E-05**	0.60
BRAF Val600Glu mutation status (mutant vs wild-type)	2.5	Before T	3.05	1.79	5.19	**4.09E-05**	1.00
		After T	0.14	0.02	1.00	* 0.05	0.65
Adjuvant chemotherapy treatment (yes vs no)	1	Before T	0.15	0.04	0.50	**2.28E-03**	0.80
		After T	0.50	0.33	0.77	**1.79E-03**	0.32

CI, confidence interval; HR, hazard ratio; PH, proportional hazard. The p-values are rounded to two decimals. *The actual p-value = 0.0505.

Supplementary Table 14. Associations between clinico-demographic/molecular variables and recurrence-free survival (RFS) in multivariable analysis.

	Cut-off time point T (year)	Time interval	HR	95% CI for HR (lower)	95% CI for HR (upper)	p-value	p-value for PH assumption test
Location (rectum vs colon)			2.43	1.45	4.07	**7.82E-04**	0.18
BRAF Val600Glu mutation status (mutant vs wild-type)	4	Before T	1.33	0.51	3.48	0.57	0.66
		After T	7.10	2.52	20.00	**2.04E-04**	0.75

CI, confidence interval; HR, hazard ratio; MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; PH, proportional hazard.

Supplementary Table 15. Associations between clinico-demographic/molecular variables and metastasis-free survival (MFS) in multivariable analysis.

	Cut-off time point T (year)	Time interval	HR	95% CI for HR (lower)	95% CI for HR (upper)	p-value	p-value for PH assumption test
Age at diagnosis			0.98	0.96	1.00	* **0.05**	0.71
Stage (II vs I)			1.92	0.95	3.86	0.07	0.27
Stage (III vs I)			3.08	1.54	6.17	**1.47E-03**	0.07
Stage (IV vs I)			1.94	0.84	4.52	0.12	0.78
Location (rectum vs colon)			1.90	1.08	3.34	**0.03**	0.80
MSI status (MSI-H vs MSI-L/MSS)			0.16	0.06	0.44	**4.45E-04**	0.27
BRAF Val600Glu mutation status (mutant vs wild-type)			3.46	2.06	5.82	**2.77E-06**	0.11
Adjuvant radiotherapy treatment (yes vs no)	5.5	Before T	0.74	0.41	1.36	0.33	0.29
		After T	6.00	1.53	23.51	**0.01**	0.86

CI, confidence interval; HR, hazard ratio; MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; PH, proportional hazard. The p-values are rounded to two decimals. *The actual p-value = 0.0496.

Supplementary Table 16. Associations between clinico-demographic/molecular variables and recurrence/metastasis-free survival (RMFS) in multivariable analysis.

	Cut-off time point T (year)	Time interval	HR	95% CI for HR (lower)	95% CI for HR (upper)	p-value	p-value for PH assumption test
Age at diagnosis			0.98	0.97	1.00	**0.03**	0.74
Stage (II vs I)			2.00	1.14	3.53	**0.02**	0.51
Stage (III vs I)			3.04	1.75	5.28	**7.66E-05**	0.26
Stage (IV vs I)			1.76	0.84	3.69	0.14	0.99
Location (rectum vs colon)	3	Before T	1.49	1.00	2.23	* 0.05	0.88
		After T	3.91	2.33	6.55	**2.34E-07**	0.76
MSI status (MSI-H vs MSI-L/MSS)			0.45	0.24	0.85	**0.01**	0.86
BRAF Val600Glu mutation status (mutant vs wild-type)			2.87	1.81	4.56	**8.12E-06**	0.63

CI, confidence interval; HR, hazard ratio; MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; PH, proportional hazard. The p-values are rounded to two decimals. *The actual p-value = 0.0524.

Supplementary Table 17. Associations between clinico-demographic/molecular variables and event-free survival (EFS) in multivariable analysis.

	Cut-off time point T (year)	Time interval	HR	95% CI for HR (lower)	95% CI for HR (upper)	p-value	p-value for PH assumption test
Stage (II vs I)			2.10	1.19	3.72	**0.01**	0.30
Stage (III vs I)	1.5	Before T	6.02	2.97	12.23	**6.61E-07**	0.46
		After T	2.99	1.57	5.71	**9.00E-04**	0.37
Stage (IV vs I)			13.18	7.56	22.96	**<2.00E-16**	0.30
Location (rectum vs colon)			1.80	1.36	2.37	**3.10E-05**	0.11
MSI status (MSI-H vs MSI-L/MSS)			0.48	0.26	0.87	**0.02**	0.67
BRAF Val600Glu mutation status (mutant vs wild-type)			2.21	1.49	3.26	**7.13E-05**	0.52
Adjuvant chemotherapy treatment (yes vs no)	1	Before T	0.40	0.22	0.72	**2.15E-03**	0.82
		After T	0.88	0.61	1.26	0.48	0.94

CI, confidence interval; HR, hazard ratio; MSI, microsatellite instability; MSI-H, microsatellite instability high; MSI-L, microsatellite instability low; MSS, microsatellite stable; PH, proportional hazard.

Appendix E: Supporting information for "A comprehensive analysis of SNPs and CNVs identifies novel markers associated with disease outcomes in colorectal cancer" (Chapter 5)

Patient cohort, clinical data, and genotype data

The NFCCR patient cohort has been described in other publications [40–42]. A total of 750 patients were collected over 5 years (1999-2003). The last follow-up date was January 2018 [42]. Clinical data was obtained from several resources, including medical charts, electronic medical records, Provincial Tumor Registry-NL/Dr. H. Bliss Murphy Cancer Centre, and Newfoundland and Labrador Center for Health Information (NLCHI) [41–43]. Microsatellite instability (MSI) status and *BRAF* Val600Glu mutation were previously identified using tumor DNAs as explained in Woods et al. [41].

The initial SNP genotype data were obtained using the Illumina® Omni1-Quad human SNP genotyping platform at an outsourced commercial facility (Centrillion Biosciences, USA) [44]. Data included 811,162 SNPs that met the following criteria: (1) SNPs that were successfully genotyped and with a missing rate ≤ 5%; (2) SNPs that satisfied the Hardy-Weinberg Equilibrium (HWE; p-value > 1×10^{-04}); (3) SNPs with minor allele counts > 2; (4) in cases when multiple SNPs shared the same genomic position, SNPs with the rs numbers were retained; and (5) SNPs that were on the autosomal chromosomes. PLINK v1.07 [28] was used to extract these data from the original datafiles. This SNP data was then used in a genetic imputation process using the software SHAPEIT (v2.r837) [45] and IMPUTE2 (v2.3.2) [46] (for details, see the following imputation section).

Imputation

Methods: The 1000 Genomes Phase 3 data (downloaded from the IMPUTE2 website: https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html) were used as the reference panel data. These data include 2,504 individuals and more than 80 million variants [47]. These individuals were individuals from different population groups, including Europeans. The IMPUTE2 developers recommend to use this inclusive reference panel because the imputation is often more accurate by using this panel than other smaller panels chosen by intuition (e.g. a panel with only Europeans) (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html). The IMPUTE2 program can automatically choose a "custom" reference panel for each individual of interest from the inclusive reference data, and this has been proved to work in variety of populations, including the homogeneous isolates (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html). The data on variants in the reference panel with 2,504 individuals were released in NCBI build 37 (hg19) coordinates, which is the same version as our genotyped SNP data.

The methodology applied in this study includes two major steps: phasing and imputation. Before phasing, genotyped SNPs were aligned to the positive DNA strand (i.e. the same strand as in the reference data). For un-ambiguous SNPs (i.e. SNPs with the allele types A/G, A/C, T/G or T/C), the strands were easy to define because the alleles would be the complementary ones if the genotyped strands were opposite of the reference strand. For example, a SNP with A/G alleles would be on the negative strand if the alleles of the same SNP in the reference data were T/C. As for the ambiguous SNPs (i.e. SNPs with alleles of A/T or C/G), similar to other studies [48,49] we made use of the MAFs and reasoned that they would be similar between our data and the data of Europeans in the reference panel. Those ambiguous SNPs with MAFs larger or equal to 40% were excluded because it is difficult to determine their strands based on the MAF. The DNA strand of the ambiguous SNPs with MAFs less than 40% were estimated by comparing their allele types to the data of Europeans in the reference data. If the minor alleles between the genotype data and the data of Europeans in the reference panel were the same, these SNPs were assumed to be on the same DNA strand. When the minor alleles were complementary to each

other, then the ambiguous SNPs in the study data were assumed to be on the negative strand; these SNPs were then flipped to the positive strand by using PLINK (v1.07) [28]. Last, SNPs with different allele types compared to the reference SNPs, and those SNPs existed in our data while not listed in the reference panel were excluded. A total of 7,244 SNPs were excluded during this step. In the end, 803,918 SNPs remained in the dataset for imputation.

The software SHAPEIT (v2.r837) [45] and IMPUTE2 (v2.3.2) [46] were used for phasing and imputation steps, respectively. Genotype data set was first separated for each chromosome using PLINK (v1.07) [28] and then phasing was performed for each chromosome as recommended in the SHAPEIT tutorial (http://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.html). During this step, the default or recommended parameters were used; --states parameter was set as its default value (100) and the effective size of 11,418 was used, which is the effective size recommended for Europeans by the developers of SHAPEIT. The same value of effective size has been used in the genetically-isolated Finland population for phasing [50]. SHAPEIT has been reported to be able to phase populations with a wide-spectrum of relatedness, including isolated populations [51].

The phased data for each chromosome were then used as the input for imputation. To do so, first, data from each chromosome were split into small segments as suggested by the tutorial provided by the IMPUTE2 program's official website (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#ex2). Imputation requires a number of genotyped SNPs/segments to construct the possible haplotypes (https://mathgen.stats.ox.ac.uk/impute/impute2_overview.html; https://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook). As recommended (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html), in this study, each chromosome was initially split into 5 Mb segments starting from the telomeres at the p-arm of each chromosome. Each segment should contain at least 200 SNPs for imputation, as suggested by other researchers (https://genome.sph.umich.edu/wiki/IMPUTE2:_1000_Genomes_Imputation_Cookbook). If this was not the case, then such segments were merged with a nearby (i.e. preceding) segment on the same chromosomal arm. Note that telomere and centromere segments may contain less than 200

SNPs as genotyping these genomic regions are problematic because of their repetitive sequences [52]. As per the segments that overlap with the centromeres, we made sure that the boundaries of the segments on the p-arm were extended to the end of each of the centromere. This also means that the start position of the next segment on the q-arm was right after the end of the centromere. If these latter segments included <200 SNPs, they were merged with the successive segment on the q-arm. The p-arms of chr 13, 14, 15, 21 and 22 did not have enough genotyped SNPs (n=4 for chr 21 and n=0 for other chrs) – so no imputation have been performed for these chromosomal arms. In the end, 548 final chromosomal segments from 22 chromosomes were generated. After this step, –int parameter was used in IMPUTE2 to conduct the imputations within each specific chromosomal segment (for example, -int 5,000,001 10,000,000 defines a segment between 5,000,001 bp and 10,000,000 bp). As for segments that were larger than 7 Mb (e.g. merged segments), an additional command -allow_large_regions was used for imputation. The parameter –Ne was set as 20,000 because IMPUTE2 developers recommend this number (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html#ex2). Other parameters were set at default values. Also, to achieve high-quality imputation for SNPs at ends of each segment, by default a buffer region of 250Kb was automatically assigned to ends of the segments.

After imputation, a number of segment-specific output files were generated for each chromosome. The data in these files were then combined together to create files (i.e. chromosome output files) that contain the imputation data per each chromosome.

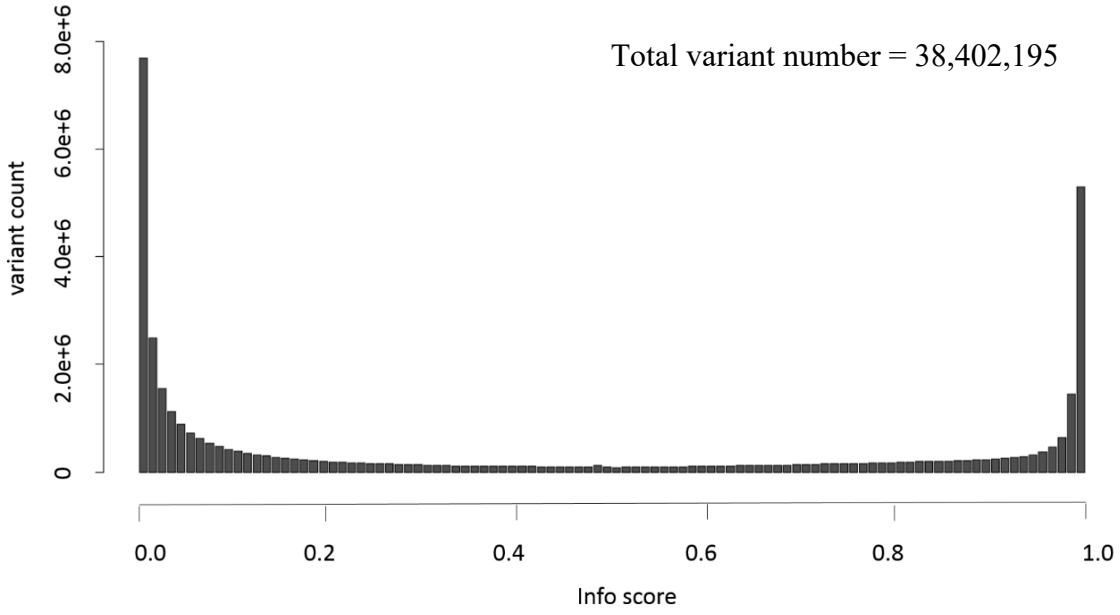
The data in the chromosome output file were then converted to PLINK PED files using GTOOL (v0.7.5) (http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html). In this process, post-QC measures were also implemented. For example, SNPs with an info score greater than 0.7 [49,53,54] and a maximum probability of the imputed genotypes larger than 90% [55] were included in the final PED files. Info score is an important indicator used to estimate imputation certainty. The closer this score is to 1.0, the higher the certainty about the imputation (http://mathgen.stats.ox.ac.uk/impute/impute_v2.html) [56]. The maximum probability of the imputed genotypes of a given SNP defines the most possible genotype of that SNP. For example, a SNP with the allele type of A/G can have three possible genotypes AA, AG and GG. After imputation, each genotype in an individual is given a "probability" value by IMPUTE2, say 0.05,

0.08 and 0.87. The maximum probability for the SNP genotype in this case is 0.87 (87%), which means the most likely genotype of the individual is GG.

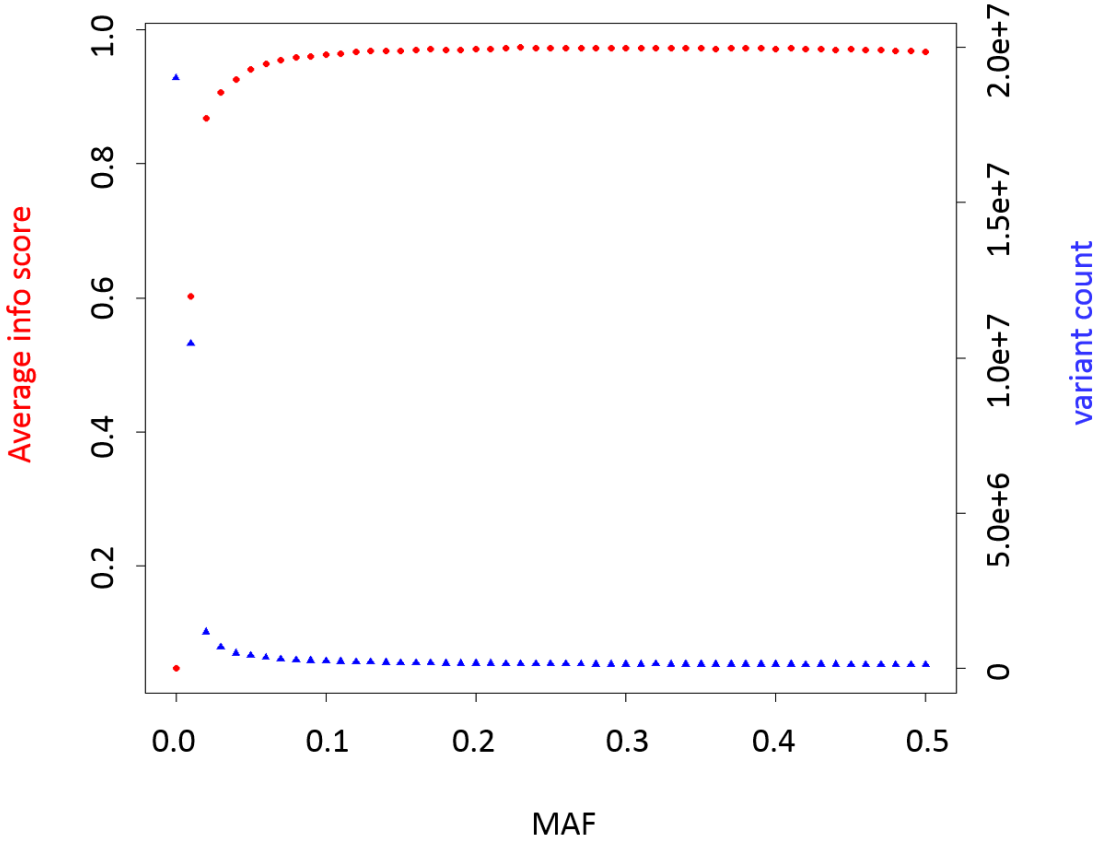
Results: More than 38 million variants were imputed with an info score > 0. The range of the concordance rate of imputations was 94% ~ 99.9% with a median of 98.7%. The concordance rate was estimated by comparing the genotypes of the known variants to their imputed genotypes and was done automatically by the IMPUTE2 program as part of its imputation process. In addition, twenty-two ambiguous SNPs that were excluded prior to phasing (one SNP per chromosome) were randomly selected and the concordance between the real and imputed genotypes were examined. The result of this examination showed that only 37 discrepancies were found among the 11,110 genotypes (22 SNPs * 505 individuals), which accounts for a concordance rate of 99.7%. Note that in the dataset the genotyped variants would have an info score and probability of 1.0. Thus, at the end the total number of variants (including genotyped ones and imputed ones) satisfying the info score and probability thresholds was 13,974,610.

The distribution of info scores for all imputed variants are shown in **Supplementary Text Figure 1**. Most variants had either very low or very high info scores. **Supplementary Text Figure 2** shows the relationship between the average info score and the MAF of the variants. Among the ~38 million imputed variants, the majority of the variants were quite rare (MAFs < 0.02) whereas ~ 6.3 million variants (~1/6) were common (i.e. had MAF ≥ 0.05) (**Supplementary Text Figure 2**). The info scores increased as the MAFs increased, as expected [49], and were particularly low for the variants with MAFs < 0.02. The average info scores for the rest of the variants (MAFs ≥ 0.02) were high (> 0.8) (**Supplementary Text Figure 2**). As shown in **Supplementary Text Figure 3**, the majority of the common SNPs (MAFs ≥ 0.05) had very high info scores, which means these variants had high imputation-quality. To be more specific about this point, 6,163,520 common and imputed variants had an info score greater than 0.7, which accounts for 97.9% of all variants with MAF ≥ 0.05. By comparing **Supplementary Text Figure 1** and **Supplementary Text Figure 3**, we can say that almost all variants with low info scores were variants with MAF < 0.05 (the bars representing the number of variants at the low

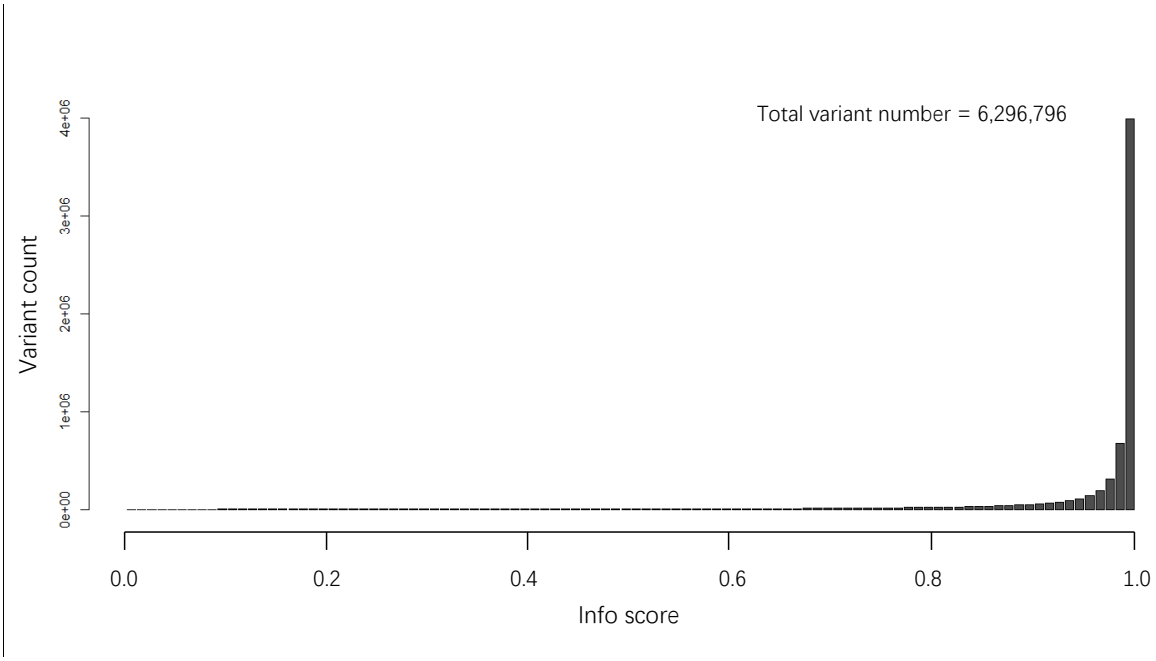
info sections of **Supplementary Text Figure 1** almost disappeared in **Supplementary Text Figure 3**). In this study, we limit our analyses to 4,711,309 SNPs that satisfied the inclusion criteria (See Methods).



Supplementary Text Figure 1. Info scores of all imputed variants. Variants were grouped by info score with 0.01 intervals. The majority of the variants had either very low or very high info scores.



Supplementary Text Figure 2. Relationship between info score and MAF. Imputed variants were grouped into 50 MAF-bins with each bin being defined as a 0.01 interval. The info scores of variants were averaged in each bin and marked with a red dot in the figure. The secondary axis indicates the number of variants in each MAF bin and the corresponding value is marked with blue triangles. The majority of the variants are with low MAFs (i.e. MAFs between 0 and 0.02) and were with low info scores (info score ≤ 0.7). Variants with higher MAFs showed high info scores, indicating that most of these variants were well-imputed SNPs.



Supplementary Text Figure 3. Info scores of the imputed variants with MAF ≥ 0.05. Variants were grouped by info score with 0.01 intervals. The majority of variants had very high info scores, indicating most of the imputed common variants were well-imputed variants.

CNV/INDEL call procedures

These analyses are described in detail in Werdyani et al. [57]. In short, MAP file and signal intensity data obtained by the Illumina® Human Omni1_Quad_v1 genome-wide SNP genotyping array (Log R ratio (LRR) and B allele frequency (BAF) measures) were used as input files to computationally predict the CNV/INDEL profiles using QuantiSNP [10] and PennCNV [9] algorithms. These algorithms are designed to detect CNVs from the whole genome SNP genotyping platform data based on a Hidden Markov Model (HMM) [9,10]. Prediction of the CNVs/INDELs by the QuantiSNP algorithm was performed using the signal intensity files of each patient using default parameters [10]. To detect the CNVs/INDELs by the PennCNV algorithm, Population Frequency of B allele (PFB) and the GC-model file for the Illumina® Human Omni1_Quad_v1 platform were generated based on the hg19 genome coordinates [9]. An adjustment of genomic waviness was implemented [11,23,58] and calls were restricted to the autosomal chromosomes [14,15].

Quality control analysis and characterization of CNVs/INDELs: Low quality CNV/INDEL calls were filtered out using the QC metrics provided by QuantiSNP and PennCNV [7,8,16,59]. We identified CNVs/INDELs that were called by both algorithms (the same copy number state (CN) and overlapped at least 50% of their sequences) using a custom Perl program [8,60]. Of note, 84.3% of such variants had identical start and end positions. In other cases, overlapping variations were merged together [7]. Since detection of CNVs/INDELs in highly repetitive sequences results in high false positive calls (e.g. centromere and telomere regions, immunoglobulin and olfactory receptor (OR) gene regions; [9,26,61]), variants that intersected at least one bp with these DNA regions were excluded from further analyses. Finally, to reduce the false-positive calls, variants that overlapped (at least 50% of their sequences) with previously experimentally validated CNVs [31,33,62] (included in the Database of Genomic Variants (DGV) [34]) were identified. These CNVs/INDELs are considered to be most likely true variations and constituted the final list of CNVs/INDELs that were predicted with high confidence. DNA analysis showed a high

concordance rate for homozygous deletions (CN state=0). For further details, please see Werdyani et al. [57].

Cut-off time point identification and inclusion in Cox models

During the process of baseline model construction, covariates that violated the PH assumption were assigned proper cut-off time points, which ensured that they satisfied the PH assumption within the time intervals defined by these cut-off time points. The method to identify the cut-off time points for variables that violate the PH assumption in Yu et al. [42] was used.

The proper cut-off time point for a given clinical variable that violated the proportional hazards (PH) assumption was identified during the backwards selection procedure, as follows: (1) time points (ranged from 0.5 years to 18.5 years, with increments of 0.5 years) were used for the variable to fit Cox models; (2) the maximized log partial likelihood values of models for each time point were obtained; and (3) the PH assumption for the variable before and after the cut-off time points in these models was checked. The proper cut-off time point was determined to be the one that makes (a) the corresponding model with the largest maximized log partial likelihood value; and (b) the PH assumption being satisfied both before and after the cut-off time point.

Variables that were not significant in the models (Cox regression p-values > 0.05) were removed one by one during the selection process. Final baseline models included significant clinical variables (Cox regression p-value < 0.05) as well as the force-entered treatment related covariates, which also satisfied the PH assumption (p-value of PH assumption test ≥ 0.05). For further details about this approach, please see Yu et al. [42].

At the time of fitting the multivariable models (i.e. when SNPs were entered into the baseline model one by one), the PH assumption was checked again for all variables in these models, including the tested genetic variants and clinical covariates. If variants violated the PH assumption, then they were analyzed in re-fitted multivariable Cox models with 5 years entered as the cut-off time point. If the covariates violated the PH assumption, then their proper cut-off

point(s) were identified/re-identified, followed by re-fitting the multivariable models as described by Yu et al. [42] (note that none of such models included variants that reached the genome-wide significance level). The final multivariable Cox models are the ones with the PH assumption satisfied for all variables.

Since the top principal component accounted for only 0.3% of the total variance in this patient cohort, principal components of genetic data were not considered as covariates.

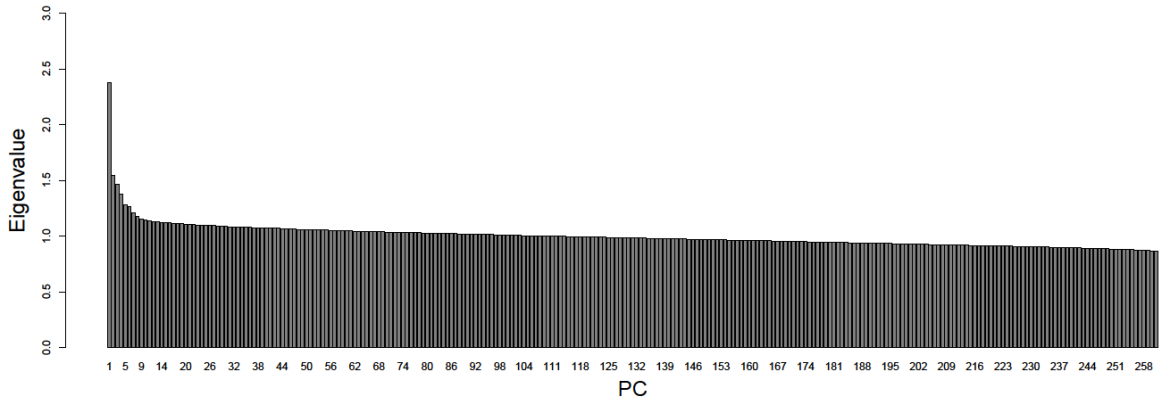
TCGA data analyses

Germline genetic data (Affymetrix genome-wide human array 6.0) of colorectal cancer patients (COAD and READ) were obtained from birdseed files (one file per patient) from the GDC Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive/search/f). SNP data from different birdseed files were combined and converted to a single plink PED/MAP file through the following steps [63]: (1) genotyping calls (in the format of allele counts 0, 1, or 2) from birdseed files were first assigned as "missing" for low-confidence SNPs (confidence value ≥ 0.1); (2) information of SNPs' genotyping calls from birdseed files were then combined; (3) probe IDs were replaced with rs numbers for all SNPs based on the information in the annotation file of the Affymetrix genome-wide human SNP array 6.0; (4) duplicated SNPs (n=2) were removed (the one with more missing data); (5) duplicated samples (n=4) were removed (the one with more missing data); (6) allele counts were converted to genotypes composed of A, T, C, and G; (7) additional required information were added to form the final PED-formatted file (sex information was derived from GDC clinical data, phenotype was assigned to 2 [i.e. affected; colorectal cancer patients]; paternal and maternal IDs were assigned to 0; Family IDs was assigned the same as Individual IDs); (8) the MAP file was constructed based on the Affymetrix annotation file. In the end, 266 patients and 906,598 SNPs were included in the PLINK PED/MAP file.

In this 266 patients cohort, patients were excluded if they (1) have any mismatched sex information (between sex information in the clinical data and the sex information imputed by

PLINK from genetic data; n=0); (2) have genotyping call rate < 5% (n=0); (3) have a high heterozygosity rate (out of 6 SD) (n=1); (4) are duplications or possible relatives (identity-by-state [IBS] PI_HAT score > 0.125) [44] (n=1); (5) are population outliers (the minimum Z score of individual's IBS distances to five nearest neighbors < -4) [28,64] (n=1); (6) are possible non-European descendants (comparing to the 1000 Genomes phase3 data in the multidimensional scaling [MDS] plot which was created based on the usage of the --genome and --mds-plot flags in PLINK1.9 [65]) (n=1). After these steps, 262 colorectal cancer patients remained in the cohort.

The genetic data of the 262 patients was then used for principal component analysis (PCA) using PLINK1.9 [65]. SNPs used for PCA were those that (1) locate on autosomal chromosomes, (2) have MAFs ≥1%, (3) have missing call rates < 5%, (4) have HWE p values ≥ 1×10^{-06}, (5) locate outside the long-LD regions [66], and (6) are independent SNPs (SNPs remained after pruning; pair-wise LD $r^2 < 0.2$)[67]. In the end, 115,051 SNPs of the 262 patients were used for PCA. The top PC (**Supplementary Text Figure 4**) accounts for 0.9% of total variance.



Supplementary Text Figure 4. Plot of Eigenvalues of principal components (PCs).

Supplementary Table 18. CNVs/INDELs examined in this study.

Variant	Variant type	Copy number status	0 copy frequency
Chr1:16152297-16153885	CNV	0, 2	0.10
Chr1:17676291-17677196	INDEL	0, 2	0.11
Chr1:58744143-58744663	INDEL	0, 2	0.76
Chr1:62082921-62083563	INDEL	0, 2	0.13
Chr1:72766413-72811692	CNV	0, 2	0.34
Chr1:80221868-80222895	CNV	0, 2	0.27
Chr1:89476427-89478432	CNV	0, 2	0.10
Chr1:92232111-92233227	CNV	0, 2	0.10
Chr1:106015878-106023356	CNV	0, 1, 2	0.14
Chr1:110187278-110188706	CNV	0, 2	0.34
Chr1:152556085-152586939	CNV	0, 2, 3	0.33
Chr1:158867802-158869733	CNV	0, 2	0.51
Chr1:159648791-159649527	INDEL	0, 2	0.56
Chr1:169207360-169241309	CNV	0, 1, 2	0.12
Chr1:179607382-179607936	INDEL	0, 2	0.13
Chr1:187717171-187722124	CNV	0, 1, 2	0.21
Chr1:194451546-194453078	CNV	0, 2	0.37
Chr1:207292578-207293178	INDEL	0, 2	0.45
Chr1:210078078-210085756	CNV	0, 2	0.37
Chr2:33224605-33227187	CNV	0, 2	0.10
Chr2:34523997-34524686	INDEL	0, 2	0.19
Chr2:34698447-34736476	CNV	0, 1, 2	0.15
Chr2:42346222-42347059	INDEL	0, 1, 2	0.19
Chr2:54565729-54567441	CNV	0, 2	0.32
Chr2:54565729-54567590	CNV	0, 1, 2	0.28
Chr2:56654397-56655541	CNV	0, 2	0.22
Chr2:70125092-70125504	INDEL	0, 2	0.15
Chr2:76773793-76775393	CNV	0, 2	0.55
Chr2:100103752-100105013	CNV	0, 2	0.20
Chr2:108855419-108856175	INDEL	0, 2	0.11
Chr2:126443281-126451762	CNV	0, 2	0.45
Chr2:127674739-127677079	CNV	0, 2	0.75
Chr2:146866008-146876881	CNV	0, 1, 2	0.21
Chr2:159959587-159961014	CNV	0, 2	0.31
Chr2:159959587-159961451	CNV	0, 2	0.18
Chr2:177268005-177271736	CNV	0, 2	0.61
Chr2:182856938-182857477	INDEL	0, 2	0.11

Chr2:194690106-194695458	CNV	0, 2	0.82
Chr2:215728845-215730688	CNV	0, 2	0.14
Chr2:227165698-227170955	CNV	0, 1, 2	0.11
Chr3:26450985-26452213	CNV	0, 2	0.22
Chr3:32102055-32106725	CNV	0, 1, 2	0.20
Chr3:47490712-47493338	CNV	0, 2	0.14
Chr3:68637537-68639440	CNV	0, 2	0.15
Chr3:68741426-68747798	CNV	0, 2	0.42
Chr3:80062959-80064447	CNV	0, 1, 2	0.37
Chr3:95465933-95468152	CNV	0, 2	0.11
Chr3:98411049-98414646	CNV	0, 2	0.15
Chr3:98900231-98902205	CNV	0, 2	0.56
Chr3:99628822-99629567	INDEL	0, 1, 2	0.10
Chr3:104278192-104279002	INDEL	0, 1, 2	0.11
Chr3:107038162-107040253	CNV	0, 2	0.10
Chr3:124936371-124936911	INDEL	0, 1, 2	0.17
Chr3:131708352-131713017	CNV	0, 1, 2	0.12
Chr3:136021052-136026101	CNV	0, 1, 2	0.13
Chr3:146387602-146390316	CNV	0, 2	0.47
Chr3:159257057-159257610	INDEL	0, 2	0.12
Chr3:162512207-162625930	CNV	0, 1, 2, 4	0.11
Chr3:162718988-162721962	CNV	0, 1, 2	0.18
Chr3:162765807-162769007	CNV	0, 2	0.52
Chr3:189737354-189740440	CNV	0, 1, 2	0.10
Chr3:192875738-192885153	CNV	0, 2, 4	0.48
Chr3:194398894-194400230	CNV	0, 2	0.36
Chr4:1602989-1603634	INDEL	0, 2	0.29
Chr4:6652161-6652800	INDEL	0, 1, 2	0.40
Chr4:6897543-6899625	CNV	0, 2	0.27
Chr4:10211268-10234260	CNV	0, 2, 3	0.49
Chr4:28421503-28422003	INDEL	0, 2	0.24
Chr4:46202844-46204937	CNV	0, 2	0.38
Chr4:61330156-61331760	CNV	0, 2	0.48
Chr4:61939506-61942200	CNV	0, 2	0.38
Chr4:64696875-64713573	CNV	0, 2	0.13
Chr4:91933043-91935779	CNV	0, 2	0.14
Chr4:115178984-115182290	CNV	0, 2, 4	0.17
Chr4:115928747-115929279	INDEL	0, 2	0.11
Chr4:133181351-133182077	INDEL	0, 2	0.13
Chr4:135433401-135435074	CNV	0, 2	0.27

Chr4:138966505-138967151	INDEL	0, 2	0.13
Chr4:142230896-142232849	CNV	0, 2	0.15
Chr4:146438871-146439991	CNV	0, 1, 2	0.16
Chr4:166003471-166004657	CNV	0, 2	0.78
Chr4:172374626-172378977	CNV	0, 2	0.11
Chr4:172989075-172992876	CNV	0, 2	0.22
Chr4:182056607-182057107	INDEL	0, 2	0.31
Chr4:186441932-186444023	CNV	0, 2	0.56
Chr4:186441932-186444110	CNV	0, 2	0.20
Chr4:187093557-187098071	CNV	0, 2	0.31
Chr5:1178511-1180425	CNV	0, 2	0.18
Chr5:1924651-1925051	INDEL	0, 2	0.68
Chr5:10273607-10274711	CNV	0, 2	0.52
Chr5:12811506-12819198	CNV	0, 1, 2	0.20
Chr5:19375544-19376421	INDEL	0, 1, 2	0.14
Chr5:21450792-21452439	CNV	0, 2	0.14
Chr5:57323612-57333211	CNV	0, 2	0.63
Chr5:60001832-60003352	CNV	0, 2	0.34
Chr5:83947987-83954795	CNV	0, 2	0.73
Chr5:83948595-83954795	CNV	0, 2	0.24
Chr5:90500630-90502051	CNV	0, 2	0.56
Chr5:97401582-97402715	CNV	0, 2	0.17
Chr5:98345182-98347056	CNV	0, 2	0.46
Chr5:106324802-106326299	CNV	0, 2	0.18
Chr5:119380452-119383902	CNV	0, 2	0.29
Chr5:135115551-135120517	CNV	0, 1, 2	0.24
Chr5:147553186-147554186	CNV	0, 2	0.33
Chr6:666535-667756	CNV	0, 2	0.39
Chr6:18402172-18402672	INDEL	0, 2	0.40
Chr6:32455482-32484368	CNV	0, 2	0.15
Chr6:32778882-32779506	INDEL	0, 2	0.20
Chr6:51736175-51736742	INDEL	0, 1, 2	0.25
Chr6:53929777-53933874	CNV	0, 2	0.47
Chr6:65347533-65349159	CNV	0, 2	0.29
Chr6:74592225-74599512	CNV	0, 1, 2	0.13
Chr6:77097876-77100253	CNV	0, 2	0.20
Chr6:77097876-77100461	CNV	0, 2	0.21
Chr6:86624320-86625771	CNV	0, 2	0.10
Chr6:89921782-89922171	INDEL	0, 2	0.32
Chr6:95193423-95194280	INDEL	0, 1, 2	0.35

Chr6:100034580-100035230	INDEL	0, 2	0.64
Chr6:114224070-114224975	INDEL	0, 2	0.11
Chr6:134269008-134269657	INDEL	0, 2	0.33
Chr6:141549108-141549906	INDEL	0, 2	0.13
Chr6:165724739-165731496	CNV	0, 2	0.21
Chr6:167488211-167489090	INDEL	0, 2	0.18
Chr7:22434643-22436658	CNV	0, 2	0.15
Chr7:24038309-24039976	CNV	0, 2	0.42
Chr7:31315876-31318783	CNV	0, 1, 2	0.23
Chr7:51594426-51598253	CNV	0, 2, 3	0.18
Chr7:62366067-62369218	CNV	0, 2	0.12
Chr7:70421545-70425749	CNV	0, 2	0.39
Chr7:73829165-73831200	CNV	0, 2	0.30
Chr7:89810608-89811996	CNV	0, 2	0.13
Chr7:89810608-89812114	CNV	0, 2	0.14
Chr7:93541865-93542465	INDEL	0, 2	0.36
Chr7:110182015-110188407	CNV	0, 1, 2	0.10
Chr7:118831427-118834466	CNV	0, 2	0.22
Chr7:126048572-126051369	CNV	0, 2	0.12
Chr7:126048572-126051476	CNV	0, 2	0.18
Chr7:133785061-133797965	CNV	0, 1, 2	0.11
Chr7:148074379-148076266	CNV	0, 2	0.67
Chr8:594761-599201	CNV	0, 2	0.68
Chr8:4122961-4124156	CNV	0, 2	0.43
Chr8:11245641-11247049	CNV	0, 2	0.18
Chr8:25066884-25070636	CNV	0, 2	0.60
Chr8:39233344-39387179	CNV	0, 1, 2	0.17
Chr8:40774744-40779338	CNV	0, 2	0.30
Chr8:42191238-42193395	CNV	0, 2	0.14
Chr8:42191238-42193599	CNV	0, 2	0.59
Chr8:75364528-75366830	CNV	0, 2	0.35
Chr8:112294125-112296209	CNV	0, 1, 2	0.38
Chr8:127192269-127194257	CNV	0, 2	0.10
Chr8:137160319-137163816	CNV	0, 2	0.26
Chr8:138742985-138743769	INDEL	0, 2	0.76
Chr9:17910043-17911627	CNV	0, 2	0.64
Chr9:22496202-22502596	CNV	0, 1, 2	0.15
Chr9:23362799-23377416	CNV	0, 1, 2	0.18
Chr9:31291381-31292431	CNV	0, 2	0.13
Chr9:71741217-71743100	CNV	0, 2	0.53

Chr9:71895369-71896250	INDEL	0, 2	0.55
Chr9:89154979-89155745	INDEL	0, 2	0.57
Chr9:101309058-101311079	CNV	0, 2	0.24
Chr9:131412549-131413853	CNV	0, 2	0.32
Chr9:131412549-131413885	CNV	0, 2	0.49
Chr9:136625265-136626037	INDEL	0, 2	0.23
Chr9:138214337-138217541	CNV	0, 2	0.16
Chr9:138479177-138480145	INDEL	0, 2	0.36
Chr10:4290129-4291584	CNV	0, 2	0.82
Chr10:4708627-4710298	CNV	0, 2	0.53
Chr10:7077551-7078246	INDEL	0, 1, 2	0.19
Chr10:27000558-27001814	CNV	0, 2	0.35
Chr10:31443722-31444976	CNV	0, 1, 2	0.22
Chr10:67306995-67314427	CNV	0, 1, 2	0.10
Chr10:78255873-78260694	CNV	0, 2	0.65
Chr10:89275888-89276407	INDEL	0, 1, 2	0.43
Chr10:93633430-93634441	CNV	0, 2	0.22
Chr10:95545536-95546273	INDEL	0, 2	0.89
Chr10:107950711-107951550	INDEL	0, 2	0.70
Chr10:114113589-114116575	CNV	0, 2	0.13
Chr10:122226947-122228534	CNV	0, 2	0.15
Chr11:5760106-5762286	CNV	0, 2	0.14
Chr11:9324025-9324496	INDEL	0, 2	0.23
Chr11:29967596-29968238	INDEL	0, 2	0.13
Chr11:31394060-31397428	CNV	0, 1, 2	0.13
Chr11:45430401-45431405	CNV	0, 2	0.30
Chr11:66712229-66713105	INDEL	0, 2	0.23
Chr11:93021163-93022144	INDEL	0, 2	0.19
Chr11:104267791-104272611	CNV	0, 2	0.49
Chr12:12026506-12026937	INDEL	0, 2	0.12
Chr12:16420184-16420943	INDEL	0, 2	0.18
Chr12:30478334-30480626	CNV	0, 1, 2	0.15
Chr12:45903118-45909531	CNV	0, 1, 2	0.29
Chr12:48709423-48710624	CNV	0, 1, 2	0.10
Chr12:60522630-60524927	CNV	0, 1, 2	0.26
Chr12:90488020-90491702	CNV	0, 2	0.18
Chr13:27050284-27052028	CNV	0, 2	0.19
Chr13:39057351-39060049	CNV	0, 2	0.43
Chr13:39934551-39935151	INDEL	0, 2	0.37
Chr13:49533568-49536464	CNV	0, 1, 2	0.12

Chr13:51069352-51072600	CNV	0, 2	0.54
Chr13:72478244-72480589	CNV	0, 1, 2	0.33
Chr13:72845850-72846775	INDEL	0, 1, 2	0.47
Chr13:90862850-90864719	CNV	0, 2	0.25
Chr13:99254450-99257146	CNV	0, 1, 2	0.10
Chr13:101894125-101896318	CNV	0, 2	0.31
Chr13:106449351-106449814	INDEL	0, 2	0.14
Chr14:20551808-20552611	INDEL	0, 2	0.29
Chr14:40615179-40617594	CNV	0, 2	0.23
Chr14:82499370-82503183	CNV	0, 2	0.38
Chr15:39372623-39373245	INDEL	0, 2	0.17
Chr15:71881673-71882625	INDEL	0, 2	0.10
Chr15:76891342-76895185	CNV	0, 1, 2	0.25
Chr15:77330786-77332606	CNV	0, 2	0.43
Chr15:91981864-91983360	CNV	0, 2	0.28
Chr16:23048233-23049446	CNV	0, 2	0.40
Chr16:48509311-48510755	CNV	0, 1, 2	0.26
Chr16:57326658-57327126	INDEL	0, 2	0.11
Chr16:58647399-58649650	CNV	0, 2	0.23
Chr16:76540062-76543447	CNV	0, 2	0.50
Chr16:78373700-78384735	CNV	0, 1, 2	0.24
Chr17:724239-724598	INDEL	0, 2	0.35
Chr17:14190726-14191429	INDEL	0, 2	0.29
Chr17:35755867-35758648	CNV	0, 1, 2	0.15
Chr17:41517334-41518185	INDEL	0, 2	0.11
Chr17:51855247-51859534	CNV	0, 1, 2	0.13
Chr17:55688120-55689796	CNV	0, 2	0.36
Chr17:76282555-76282929	INDEL	0, 2	0.28
Chr18:5324676-5326221	CNV	0, 2	0.27
Chr18:24571673-24572190	INDEL	0, 2	0.17
Chr18:35306101-35306609	INDEL	0, 2	0.49
Chr18:38862147-38868004	CNV	0, 2	0.60
Chr18:38864903-38868004	CNV	0, 2	0.23
Chr18:47695103-47698268	CNV	0, 2	0.11
Chr18:54946766-54948517	CNV	0, 2	0.56
Chr18:63766950-63769066	CNV	0, 2	0.11
Chr18:75267039-75267968	INDEL	0, 2	0.33
Chr18:77310162-77312078	CNV	0, 1, 2	0.42
Chr19:2909643-2910369	INDEL	0, 2	0.79
Chr19:5510301-5510667	INDEL	0, 2	0.11

Chr19:12694963-12697389	CNV	0, 2	0.48
Chr19:13776129-13776658	INDEL	0, 2	0.12
Chr19:15046722-15047605	INDEL	0, 2	0.28
Chr19:31287833-31289043	CNV	0, 2	0.30
Chr19:51406972-51407935	INDEL	0, 2	0.11
Chr20:1389773-1390682	INDEL	0, 2	0.12
Chr20:1389773-1391436	CNV	0, 2	0.14
Chr21:16588414-16589135	INDEL	0, 2	0.19
Chr21:19327135-19328810	CNV	0, 1, 2	0.11
Chr21:44970373-44973184	CNV	0, 2	0.30
Chr21:47388151-47389593	CNV	0, 2	0.28
Chr22:18058001-18059664	CNV	0, 2	0.80
Chr22:24274775-24276797	CNV	0, 2	0.17
Chr22:24365041-24367511	CNV	0, 2	0.14
Chr22:35645524-35646052	INDEL	0, 2	0.43
Chr22:37143405-37146870	CNV	0, 2	0.50
Chr22:39295546-39298533	CNV	0, 1, 2	0.18

CNV, copy number variation; INDEL, insertion/deletion. Data based on a previous study [57] of our group.

Supplementary Table 19. Pair-wise Pearson correlation coefficients of clinico-demographic variables, MSI status, and *BRAF* Val600Glu mutation in the SNP analysis cohort with 505 patients.

	Sex	Grade	Histology	Location	Stage	MSI status	*BRAF* Val600Glu mutation	Adjuvant chemotherapy treatment	Adjuvant radiotherapy treatment
Sex	1.00	0.02	-0.09	0.09	0.04	-0.12	-0.19	0.08	0.06
Grade		1.00	0.07	0.01	0.08	0.13	0.10	0.05	0.05
Histology			1.00	-0.12	0.09	0.09	0.08	0.03	-0.09
Location				1.00	-0.04	-0.19	-0.22	0.22	0.69
Stage					1.00	-0.09	0.04	0.40	0.14
MSI status						1.00	0.39	-0.04	-0.11
BRAF Val600Glu mutation							1.00	0.02	-0.14
Adjuvant chemotherapy treatment								1.00	0.51
Adjuvant radiotherapy treatment									1.00

MSI, microsatellite instability.

Supplementary Table 20. The number of genetic variants analyzed in the univariate and multivariable analyses.

SNPs						
	Additive model		**Dominant model**		**Recessive model**	
	Number of variants	*Number of variants entered into the multivariable analysis	Number of variants	*Number of variants entered into the multivariable analysis	Number of variants	*Number of variants entered into the multivariable analysis
DSS - univariate						
Satisfying the PH assumption (No cut-off time)	4,464,856	15	4,470,070	1	4,507,677	271
Satisfying the PH assumption using 5 years as the cut-off time	244,790	10	239,574	2	203,103	149
**Still violating PH assumption after using the 5 years as the cut-off time	1,663	NA	1,665	NA	529	NA
RMFS - univariate						
Satisfying the PH assumption (No cut-off time)	4,471,130	152	4,470,972	3	4,497,910	338
Satisfying the PH assumption using 5 years as the cut-off time	179,755	17	181,300	0	169,358	315
**Still violating the PH assumption after using 5 years as the cut-off time	60,424	NA	59,037	NA	44,041	NA

CNVs/INDELs		
	Number of variants	*Number of variants entered into the multivariable analysis
DSS - univariate		
Satisfying the PH assumption (No cut-off time)	235	0
Satisfying the PH assumption using 5 years as the cut-off time	19	0
**Still violating the PH assumption after using 5 years as the cut-off time	0	NA
RMFS - univariate		
Satisfying the PH assumption (No cut-off time)	243	0
Satisfying the PH assumption using 5 years as the cut-off time	10	0

**Still violating the PH assumption after using 5 years as the cut-off time	1	NA

CI, confidence interval; CNV, copy number variation; DSS, disease-specific survival; INDEL, insertion/deletion; NA, not applicable; PH, proportional hazards; RMFS, recurrence/metastasis-free survival; SNP, single nucleotide polymorphism.

*, the univariate p value threshold for variants to enter into the multivariable analysis is 5×10^{-06}. **, excluded from further analysis. Note that SNPs that passed the univariate p value threshold were not entered into multivariable analysis if their upper limits of the 95% CIs were infinity.

Supplementary Table 21. Baseline characteristics of the TCGA colorectal cancer patient cohort.

Variable	Number (n=262 in total)	%
Tumor location		
Colon	188	71.76
Rectum	74	28.24
Stage		
I	41	15.65
II	91	34.73
III	82	31.30
IV	38	14.50
Unknown	10	3.82
MSI		
MSI-L/MSS	217	82.82
MSI-H	44	16.79
Unknown	1	0.38
Follow-up time		
Median (range)	2.04 (0 - 11.70)	-
DSS status		
Death from other causes or alive	218	83.21
Death from colorectal cancer	24	9.16
Unknown	20	7.63

DSS, disease-free survival; MSI, microsatellite instability; MSI-H, microsatellite instability-high; MSI-L, microsatellite instability-low; MSS, microsatellite stable. Data based on the GDC (https://portal.gdc.cancer.gov/) and a study [68] published in 2018.

Supplementary Table 22. SNPs identified to be significantly associated with disease-specific survival (DSS) in multivariable analysis under the *recessive* genetic model.

Variant	Chr	Position	Minor/ major allele	MAF	Variant type	Info Score	Time period post-diagnosis	#HR (95% CI)	p value	p value of the PH assump-tion test	*Located region
rs28552674	1	235583778	T/C	0.10	Imputed	0.998	-	44.16 (12.04, 161.98)	**1.12×10^{-08}**	0.89	Intron of *TBCE*
rs12758637	1	235584193	A/G	0.10	Imputed	1	-	44.16 (12.04, 161.98)	**1.12×10^{-08}**	0.89	Intron of *TBCE*
rs11579933	1	235588409	A/C	0.10	Imputed	1	-	44.16 (12.04, 161.98)	**1.12×10^{-08}**	0.89	Intron of *TBCE*
kgp2690683	1	235590559	A/G	0.10	Genotyped	-	-	44.16 (12.04, 161.98)	**1.12×10^{-08}**	0.89	Intron of *TBCE*
rs71640701	1	235608749	A/G	0.10	Imputed	0.998	-	44.16 (12.04, 161.98)	**1.12×10^{-08}**	0.89	Intron of *TBCE*
rs72239609	1	235609498	CT/C	0.10	Imputed	0.983	-	45.89 (12.50, 168.44)	**8.03×10^{-09}**	0.85	Intron of *TBCE*
rs6429094	1	235611093	G/A	0.11	Genotyped	-	-	44.16 (12.04, 161.98)	**1.12×10^{-08}**	0.89	Intron of *TBCE,* 3' UTR of *B3GALNT2*
rs35242859	1	235612394	CAGT T/C	0.10	Imputed	1	-	44.16 (12.04, 161.98)	**1.12×10^{-08}**	0.89	3' UTR of *B3GALNT2*
rs358373	3	13188055	G/A	0.12	Imputed	0.970	-	17.87 (6.45, 49.49)	**2.88×10^{-08}**	0.44	5' of *IQSEC1*
rs530425	3	13189919	A/G	0.12	Imputed	0.986	-	17.36 (6.31, 47.76)	**3.28×10^{-08}**	0.59	5' of *IQSEC1*
rs140970549	3	164282763	T/TTTC	0.17	Imputed	1	-	9.35 (4.33, 20.22)	**1.31×10^{-08}**	0.93	3' of *SI*
rs58844954	3	164330365	C/T	0.17	Imputed	0.995	-	9.32 (4.31, 20.15)	**1.38×10^{-08}**	0.93	3' of *SI*

Chr, chromosome; CI, confidence interval; HR, hazard ratio; MAF, minor allele frequency; PH, proportional hazards. #, Hazard ratio was estimated under the recessive genetic model for AA vs [AB+BB], where A is the minor allele and B is the major allele. *, Gene annotation is obtained from the UCSC database ("UCSC genes" from the UCSC browser [GRCh37/hg19]) [13], and only the overlapped (for SNPs within genes) or the closest (for SNPs in Intergenic regions) genes are shown in this Table. 3', downstream of the gene. 5', upstream of the gene. Models are adjusted for MSI status, disease stage, tumor location (6 years as the cut-off time point), adjuvant chemotherapy and radiotherapy statuses (7 years as the cut-off time point for adjuvant radiotherapy). SNPs that are in high-LD ($r^2 > 0.8$) with each other on the same chromosome are highlighted.

Supplementary Table 23. SNPs identified to be significantly associated with recurrence/metastasis-free survival (RMFS) in multivariable analysis under the *recessive* model.

Variant	Chr	Position	Minor/ major allele	MAF	Variant type	Info score	Time period post-diagno-sis	#HR (95% CI)	p value	p value of the PH assump-tion test	*Located region
rs4534237	2	114160484	G/C	0.12	Imputed	0.975	-	9.69 (4.50, 20.85)	**6.27×10^{-09}**	0.49	5' of *CBWD2*
rs75261537	2	114162255	C/A	0.12	Imputed	0.975	-	9.69 (4.50, 20.85)	**6.27×10^{-09}**	0.49	5' of *CBWD2*
rs72641537	4	64327395	T/C	0.10	Imputed	0.982	-	20.18 (6.91, 58.9)	**3.86×10^{-08}**	0.77	3' of *TECRL*
rs11307057	4	94764199	T/TG	0.16	Imputed	0.98	-	8.29 (3.91, 17.58)	**3.49×10^{-08}**	0.52	3' of *ATOH1*
rs11193950	10	109966710	T/C	0.14	Imputed	0.993	-	20.54 (8.47, 49.82)	**2.33×10^{-11}**	0.54	5' of *SORCS1*
rs10884600	10	109969396	G/A	0.14	Genotyped	-	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193953	10	109971822	C/T	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193956	10	109972622	A/G	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs144602401	10	109973954	C/CTATT	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193958	10	109974184	T/C	0.15	Imputed	0.973	-	17.08 (7.49, 38.96)	**1.52×10^{-11}**	0.45	5' of *SORCS1*
rs113965753	10	109975195	A/AG	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193959	10	109979273	C/T	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193962	10	109983961	A/G	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs12266409	10	109986116	G/A	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*

rs12357671	10	109987906	T/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs10509868	10	109988645	G/A	0.14	Genotyped	-	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs12355000	10	109988766	C/A	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193967	10	109996302	C/T	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193968	10	109996688	T/A	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193971	10	109998171	A/T	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193975	10	109999737	A/G	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193976	10	110003244	T/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs7920125	10	110003692	T/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs7920134	10	110003720	G/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193977	10	110004373	A/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193978	10	110004533	A/G	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs58143932	10	110005552	T/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193980	10	110006462	A/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs112377485	10	110016099	G/GA	0.15	Imputed	0.991	-	20.6 (8.49, 49.99)	**2.23×10^{-11}**	0.55	5' of *SORCS1*
rs17124169	10	110017139	C/A	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193989	10	110021384	T/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*
rs11193990	10	110021388	T/C	0.14	Imputed	1	-	20.85 (8.59, 50.56)	**1.84×10^{-11}**	0.54	5' of *SORCS1*

rs9988666	10	110032277	C/T	0.14	Imputed	0.996	-	20.72 (8.54, 50.25)	**1.99×10^{-11}**	0.55	5' of *SORCS1*
rs10884603	10	110033909	A/C	0.14	Imputed	0.995	-	20.72 (8.54, 50.25)	**1.99×10^{-11}**	0.55	5' of *SORCS1*
rs10884604	10	110035202	G/A	0.14	Imputed	0.994	-	20.72 (8.54, 50.25)	**1.99×10^{-11}**	0.55	5' of *SORCS1*
rs10884605	10	110035573	A/T	0.14	Imputed	0.994	-	20.72 (8.54, 50.25)	**1.99×10^{-11}**	0.55	5' of *SORCS1*
rs10884606	10	110037946	A/C	0.14	Imputed	0.993	-	20.72 (8.54, 50.25)	**1.99×10^{-11}**	0.55	5' of *SORCS1*
rs117213049	10	110045564	T/A	0.14	Imputed	0.972	-	21.24 (8.74, 51.63)	**1.54×10^{-11}**	0.57	5' of *SORCS1*
rs77206409	11	19420899	A/C	0.13	Imputed	0.991	-	34.91 (13.11, 92.97)	**1.17×10^{-12}**	0.79	Intron of *NAV2*
rs79632817	11	19426141	C/G	0.12	Imputed	0.972	-	35.59 (13.32, 95.15)	**1.08×10^{-12}**	0.80	Intron of *NAV2*
rs78348500	11	19427476	G/T	0.12	Imputed	0.967	-	35.59 (13.32, 95.15)	**1.08×10^{-12}**	0.80	Intron of *NAV2*
rs76656890	11	19427968	T/G	0.12	Imputed	0.963	-	35.55 (13.30, 95.03)	**1.09×10^{-12}**	0.80	Intron of *NAV2*
rs12225106	11	86443409	A/G	0.11	Imputed	0.995	-	113.55 (23.37, 551.59)	**4.41×10^{-09}**	0.89	5’ of *ME3*
rs10734945	12	128295046	C/T	0.13	Imputed	0.992	-	15.92 (6.17, 41.11)	**1.07×10^{-08}**	0.79	5’ of *FLJ37505*
rs755836	12	128297785	T/C	0.13	Genotyped	-	-	16.02 (6.21, 41.38)	**9.97×10^{-09}**	0.79	5’ of *FLJ37505*
rs1882266	12	128298247	C/T	0.13	Imputed	0.998	-	16.02 (6.21, 41.38)	**9.97×10^{-09}**	0.79	5’ of *FLJ37505*
rs1653376	12	128301795	C/G	0.13	Imputed	0.994	-	16.07 (6.22, 41.52)	**9.84×10^{-09}**	0.81	5’ of *FLJ37505*
rs1653375	12	128301798	T/C	0.13	Imputed	0.994	-	16.07 (6.22, 41.52)	**9.84×10^{-09}**	0.81	5’ of *FLJ37505*
rs11348435	12	128303354	GA/G	0.12	Imputed	0.955	-	33.27 (9.67, 114.49)	**2.73×10^{-08}**	0.88	5’ of *FLJ37505*
rs7206003	16	75205764	A/G	0.10	Genotyped	-	-	95.42 (20.14, 452.01)	**9.26×10^{-09}**	0.87	3’ UTR of *ZFP1*

rs12922107	16	75206392	G/C	0.10	Imputed	0.996	-	94.6 (19.97, 448.12)	**9.88×10^{-09}**	0.87	3' of *ZFP1*
rs12923789	16	75206439	A/G	0.10	Imputed	0.996	-	94.6 (19.97, 448.12)	**9.88×10^{-09}**	0.87	3' of *ZFP1*
rs7188765	16	75208536	G/A	0.10	Imputed	0.962	-	93.19 (19.68, 441.37)	**1.10×10^{-08}**	0.87	3' of *ZFP1*
rs3064467	17	63385829	TAAA C/T	0.11	Imputed	0.950	-	70.27 (17.35, 284.66)	**2.56×10^{-09}**	0.88	3' of *AXIN2*
rs6507174	18	34121268	A/G	0.15	Imputed	0.995	-	9.34 (4.51, 19.35)	**1.85×10^{-09}**	0.45	Intron of *FHOD3*
rs1874381	18	34122410	A/G	0.15	Genotyped	-	-	9.33 (4.50, 19.33)	**1.88×10^{-09}**	0.46	Intron of *FHOD3*
rs817090	2	49575028	T/C	0.12	Imputed	0.985	Before 5 years	0.50 (0.07, 3.57)	4.87×10^{-01}	0.68	5' of *FSHR*
							After 5 years	40.45 (11.57, 141.49)	**6.96×10^{-09}**	0.80	
rs200143895	12	119707968	A/AAAAG	0.12	Imputed	0.986	Before 5 years	3.87 (0.92, 16.24)	6.42×10^{-02}	0.46	3' of *LINC00934*
							After 5 years	141.03 (27.74, 716.86)	**2.44×10^{-09}**	0.98	
rs11064732	12	119709787	T/A	0.12	Imputed	0.995	Before 5 years	3.89 (0.93, 16.29)	6.33×10^{-02}	0.48	3' of *LINC00934*
							After 5 years	135.71 (26.87, 685.51)	**2.81×10^{-09}**	0.97	

Chr, chromosome; CI, confidence interval; HR, hazard ratio; MAF, minor allele frequency; PH, proportional hazards. #, Hazard ratio was estimated under the recessive genetic model for AA vs [AB+BB], where A is the minor allele and B is the major allele. *, Gene annotation is derived from the UCSC database ("UCSC genes" from the UCSC browser [GRCh37/hg19]) [13], and only the overlapped (for SNPs within genes) or the closest (for SNPs in intergenic regions) genes are shown in this Table. 3', downstream of the gene. 5', upstream of the gene. Models are adjusted for disease stage, tumor location (3 years as the cut-off time point), adjuvant chemotherapy and radiotherapy statuses. SNPs that are in high-LD ($r^2 > 0.8$) with each other on the same chromosome are highlighted.

Supplementary Table 24. Top SNPs in multivariable analysis that have nominal/suggestive associations with recurrence/metastasis-free survival (RMFS) under the dominant and additive genetic models.

Genetic model	Variant	Chr	Position	Minor/ major allele	MAF	Variant type	Info score	Time period post-diagnosis	#HR (95% CI)	p value	p value of the PH assump-tion test
Dominant											
	rs1372330	9	119519588	A/G	0.14	Genotyped	-	-	2.11 (1.45, 3.05)	8.06×10^{-05}	0.22
	rs979746	17	46336112	A/C	0.19	Imputed	0.994	-	2.32 (1.61, 3.35)	7.30×10^{-06}	0.50
	rs73151111	21	25723030	G/A	0.14	Imputed	0.969	-	2.27 (1.57, 3.30)	1.43×10^{-05}	0.07
Additive											
	*rs71011025	2	185087530	AT/A	0.18	Imputed	0.989	-	2.22 (1.66, 2.98)	1.01×10^{-07}	0.20
	*rs13400857	2	185203547	G/T	0.19	Imputed	0.979	-	2.18 (1.63, 2.92)	1.84×10^{-07}	0.15
	*rs34039920	2	185227401	T/TAA	0.19	Imputed	0.997	-	2.22 (1.64, 2.99)	1.82×10^{-07}	0.20
	rs10160322	11	107683902	A/G	0.11	Imputed	0.977	Before 5 years	2.45 (1.70, 3.54)	1.45×10^{-06}	0.10
								After 5 years	1.17 (0.36, 3.86)	7.93×10^{-01}	0.97
	rs10160657	11	107683926	T/A	0.11	Imputed	0.985	Before 5 years	2.45 (1.71, 3.51)	1.03×10^{-06}	0.08
								After 5 years	1.08 (0.33, 3.59)	8.98×10^{-01}	0.96
	rs12808659	11	107685038	G/A	0.11	Imputed	0.989	Before 5 years	2.41 (1.68, 3.47)	1.83×10^{-06}	0.07
								After 5 years	1.02 (0.31, 3.41)	9.72×10^{-01}	0.96

Chr, chromosome; CI, confidence interval; HR, hazard ratio; MAF, minor allele frequency; PH, proportional hazards; RMFS, recurrence/metastasis-free survival. #, Hazard ratio was estimated under the dominant genetic model for [AA+AB] vs BB and under the additive genetic model for AA vs AB vs

BB, where A is the minor allele and B is the major allele. *, Note that rs13400857 is in high-LD ($r^2 > 0.8$) with the other two SNPs (rs71011025 and rs34039920), but rs71011025 and rs34039920 are not in high-LD with each other ($r^2 = 0.77$). Models are adjusted for disease stage, tumor location (3 years as the cut-off time point), adjuvant chemotherapy and radiotherapy statuses. For the additive genetic model, results shown include the top three SNPs both with and without the cut-off time point of 5 years. SNPs that are in high-LD ($r^2 > 0.8$) with each other on the same chromosome are highlighted.

Supplementary Table 25. eQTLs (identified and high-LD variants) in DSS and RMFS recessive models.

Outcome - genetic model	rs ID	*eQTL associated gene (tissue) - RegulomeDB	*eQTL associated gene (tissue) - GTEx	High-LD SNP	SNP(s) identified in our study
DSS-recessive	#rs12757197 (kgp2690683)	*TBCE* (monocyte)	*TBCE* (transverse colon)	No	rs12757197 (kgp2690683)
DSS-recessive	rs28552674	-	*TBCE* (transverse colon)	No	rs28552674
DSS-recessive	rs12758637	-	*TBCE* (transverse colon)	No	rs12758637
DSS-recessive	rs11579933	-	*TBCE* (transverse colon)	No	rs11579933
DSS-recessive	rs71640701	-	*TBCE* (transverse colon)	No	rs71640701
DSS-recessive	rs6429094	-	*TBCE* (transverse colon)	No	rs6429094
DSS-recessive	rs35242859	-	*TBCE* (transverse colon)	No	rs35242859
DSS-recessive	rs7412979	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701
DSS-recessive	rs12726892	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701
DSS-recessive	rs7537	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701
DSS-recessive	rs34729832	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701

DSS-recessive	rs6702967	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701
DSS-recessive	rs36073314	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701
DSS-recessive	rs12087848	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701
DSS-recessive	rs6696235	-	*TBCE* (transverse colon)	Yes	rs11579933, kgp2690683, rs12758637, rs28552674, rs35242859, rs6429094, and rs71640701
RMFS-recessive	rs7188765	-	*ZFP1* (transverse colon)	No	rs7188765
RMFS-recessive	rs12716782	-	*ZFP1* (transverse colon)	Yes	rs7206003, rs12922107, rs12923789, and rs7188765
RMFS-recessive	rs7189541	-	*ZFP1* (transverse colon)	Yes	rs7206003, rs12922107, rs12923789, and rs7188765
RMFS-recessive	rs11648915	-	*ZFP1* (transverse colon)	Yes	rs7206003, rs12922107, rs12923789, and rs7188765
RMFS-recessive	rs6564214	-	*ZFP1* (transverse colon)	Yes	rs7206003, rs12922107, rs12923789, and rs7188765
RMFS-recessive	rs9931007	-	*ZFP1* (transverse colon)	Yes	rs7206003, rs12922107, rs12923789, and rs7188765

DSS, disease-specific survival; eQTL, expression quantitative trait locus; LD, linkage disequilibrium; RMFS, recurrence/metastasis-free survival; SNP, single nucleotide polymorphism. *****, all SNPs identified in recessive models as well as those SNPs that are in high-LD with them (retrieved from Haploreg [69]) were explored in RegulomeDB [70] and GTEx [71]. Note that GTEx data are shown for colon tissue, as it has no data for rectal tissue. [#], the rs number of the identified SNP

kgp2690683 is rs12757197; rs number was identified by SNP's genomic positions and alleles. The eQTLs are all cis-eQTLs that locate within ±1 Mb of the transcription start sites of the genes shown in the Table.

Supplementary Table 26. Association between *WBP11* expression levels and consensus molecular subtypes (CMS).

	P value of Kruskal-Wallis test	P value of Dunn's test for pair-wise comparison		
		Pairs	P value	Adjusted p value (Bonferroni method)
CMS (1, 2, 3, 4)	**9.66×10^{-07}**	CMS2 vs CMS1	**2.96×10^{-03}**	**1.77×10^{-02}**
		CMS3 vs CMS1	7.51×10^{-02}	4.51×10^{-01}
		CMS4 vs CMS1	**7.30×10^{-08}**	**4.38×10^{-07}**
		CMS3 vs CMS2	6.32×10^{-01}	1.00
		CMS4 vs CMS2	**1.22×10^{-03}**	**7.35×10^{-03}**
		CMS4 vs CMS3	**9.02×10^{-03}**	5.41×10^{-02}

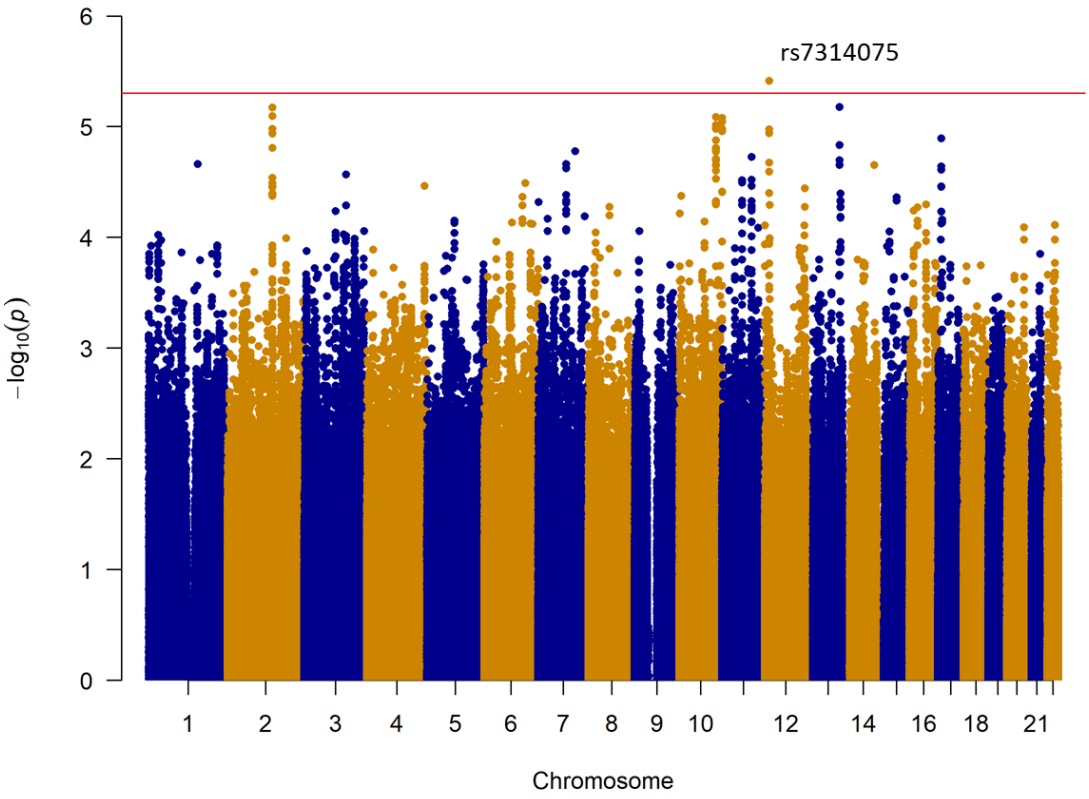
CMS, Consensus Molecular Subtypes.

Supplementary Table 27. Top CNVs/INDELs in univariate analysis of the disease-specific survival (DSS) and recurrence/metastasis-free survival (RMFS).

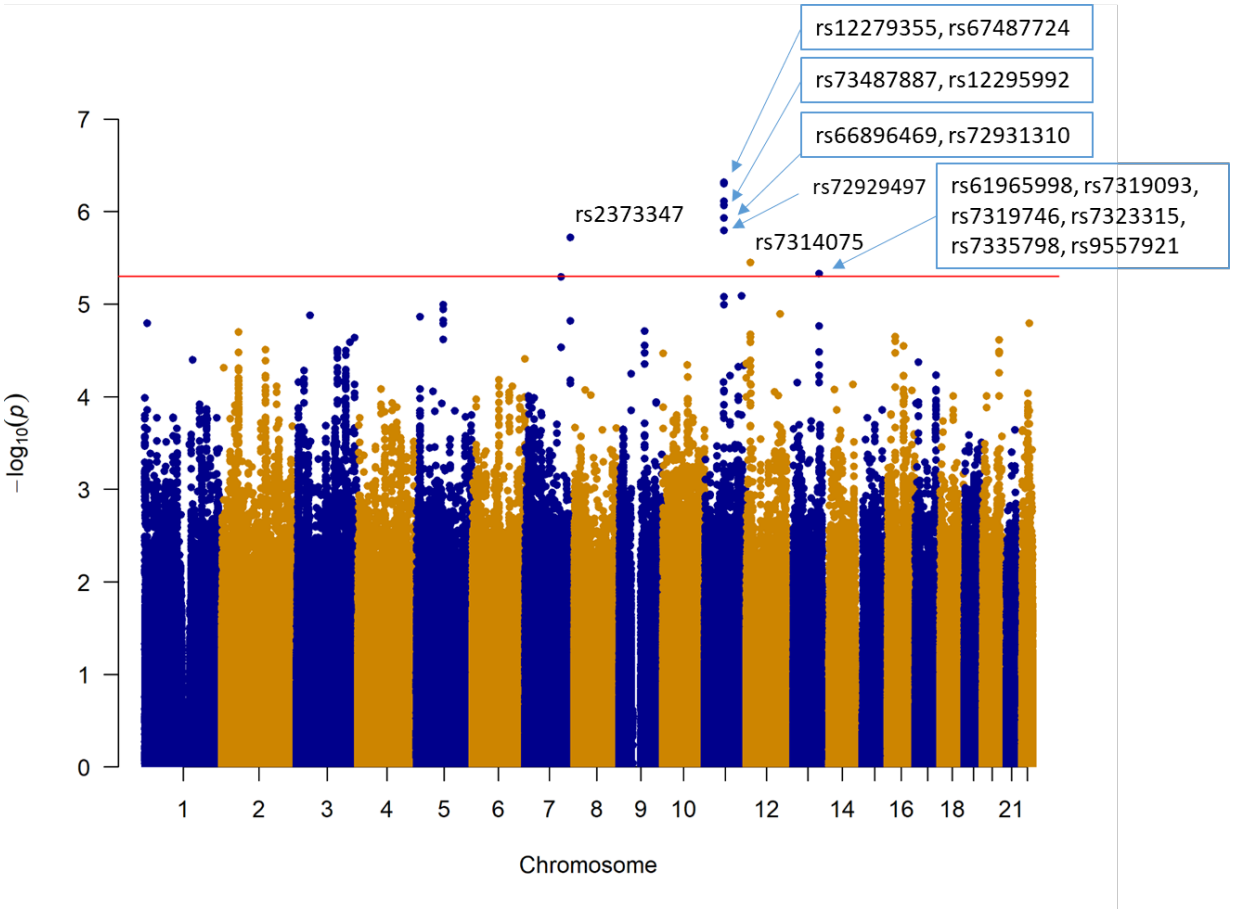
Outcome	Variant	0 copy frequency	Variant type	Time period post-diagnosis	#HR (95% CI)	p value	p value of the PH assumption test
DSS							
	Chr2:227165698-227170955 (0 copy vs 1 or 2 copies)	0.11	CNV	-	2.28 (1.40, 3.73)	9.91×10^{-4}	0.73
	Chr7:73829165-73831200 (0 copy vs 2 copies)	0.30	CNV	-	1.88 (1.26, 2.80)	1.95×10^{-3}	0.10
	Chr19:15046722-15047605 (0 copy vs 2 copies)	0.28	INDEL	-	0.45 (0.26, 0.76)	3.13×10^{-3}	0.30
	Chr2:54565729-54567441 (0 copy vs 2 copies)	0.32	CNV	Before 5 years	0.68 (0.38, 1.20)	1.84×10^{-01}	1.00
				After 5 years	2.37 (1.24, 4.52)	9.02×10^{-03}	0.58
	Chr2:76773793-76775393 (0 copy vs 2 copies)	0.55	CNV	Before 5 years	0.86 (0.52, 1.42)	5.64×10^{-01}	1.00
				After 5 years	3.56 (1.56, 8.10)	2.51×10^{-03}	0.54
	Chr9:22496202-22502596 (0 copy vs 1 or 2 copies)	0.15	CNV	Before 5 years	0.73 (0.33, 1.60)	4.30×10^{-01}	0.73
				After 5 years	2.693 (1.33, 5.45)	5.94×10^{-03}	0.69
RMFS							
	Chr1:62082921-62083563 (0 copy vs 2 copies)	0.13	INDEL	-	1.97 (1.25, 3.10)	3.47×10^{-03}	0.48
	Chr2:146866008-146876881 (0 copy vs 1 or 2 copies)	0.21	CNV	-	0.48 (0.28, 0.813)	6.30×10^{-03}	0.69

	Chr7:24038309-24039976 (0 copy vs 2 copies)	0.42	CNV	-	0.58 (0.39, 0.85)	5.37×10^{-03}	0.90
	Chr2:76773793-76775393 (0 copy vs 2 copies)	0.55	CNV	Before 5 years	1.04 (0.71, 1.53)	8.34×10^{-01}	0.15
				After 5 years	2.91 (0.95, 8.92)	6.19×10^{-02}	0.89
	Chr4:172374626-172378977 (0 copy vs 2 copies)	0.11	CNV	Before 5 years	0.53 (0.25, 1.14)	1.05×10^{-01}	0.06
				After 5 years	0.94 (0.21, 4.10)	9.32×10^{-01}	0.93
	Chr22:35645524-35646052 (0 copy vs 2 copies)	0.43	INDEL	Before 5 years	0.61 (0.40, 0.91)	1.61×10^{-02}	0.19
				After 5 years	2.18 (0.81, 5.90)	1.24×10^{-01}	0.93

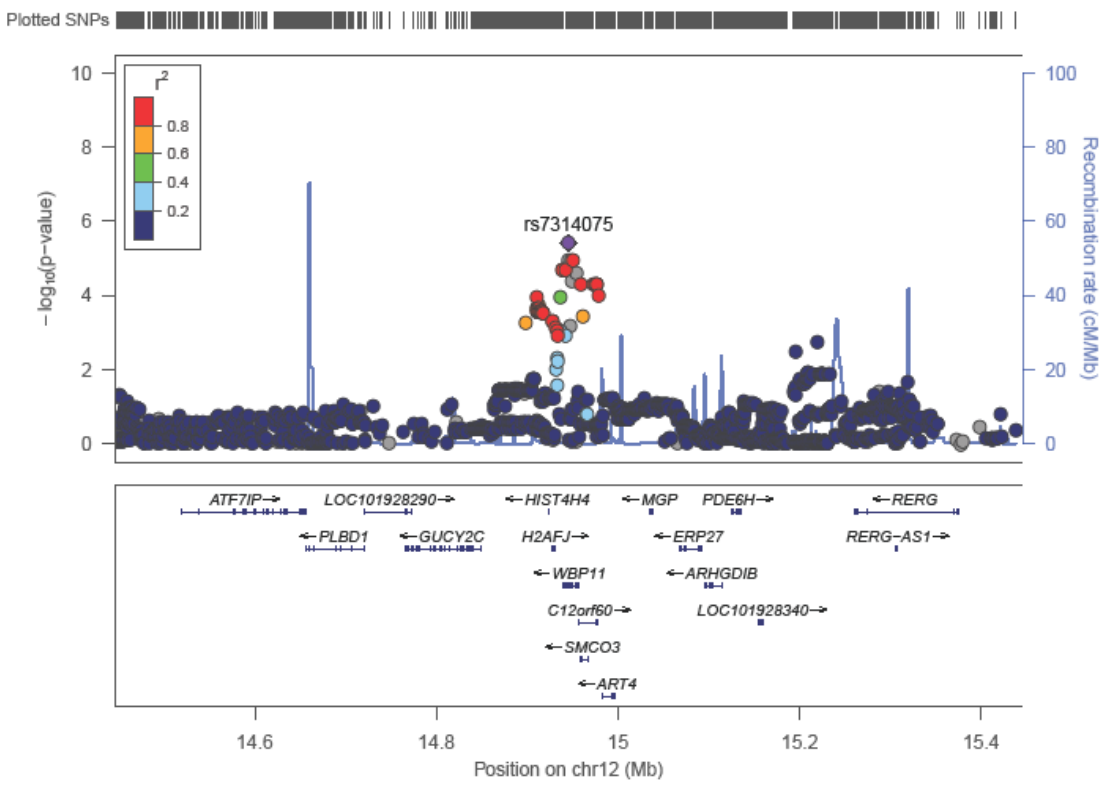
CI, confidence interval; CNV, copy number variation; DSS, disease-specific survival; HR, hazard ratio; INDEL, insertion/deletion; PH, proportional hazards; RMFS, recurrence/metastasis-free survival. [#], Hazard ratio was estimated for 0 copy vs at least one copy. For DSS and RMFS analyses, results shown include the top three CNVs/INDELs both with and without the cut-off time point of 5 years.



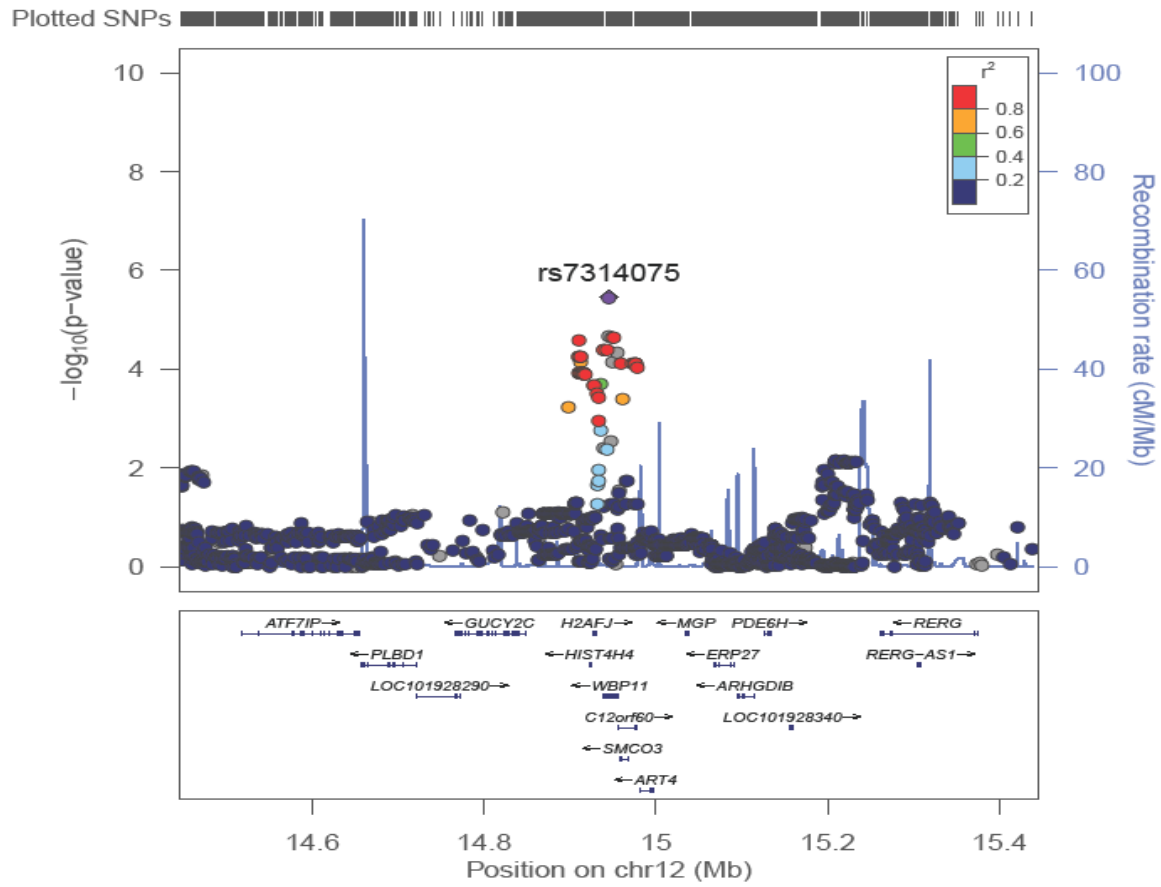
Supplementary Figure 7. Manhattan plot showing the SNP (i.e. rs7314075) with a p value that passed the 5×10^{-06} threshold (indicated by the red line) in the univariate Cox regression analysis (DSS; *dominant* genetic model). Manhattan plot was generated using SNPs that satisfied the PH assumption in the univariate analysis. Note the SNP that is indicated in this figure is the SNP that passed the significance level of 5×10^{-08} in the multivariable analysis (see **Table 5.2**).



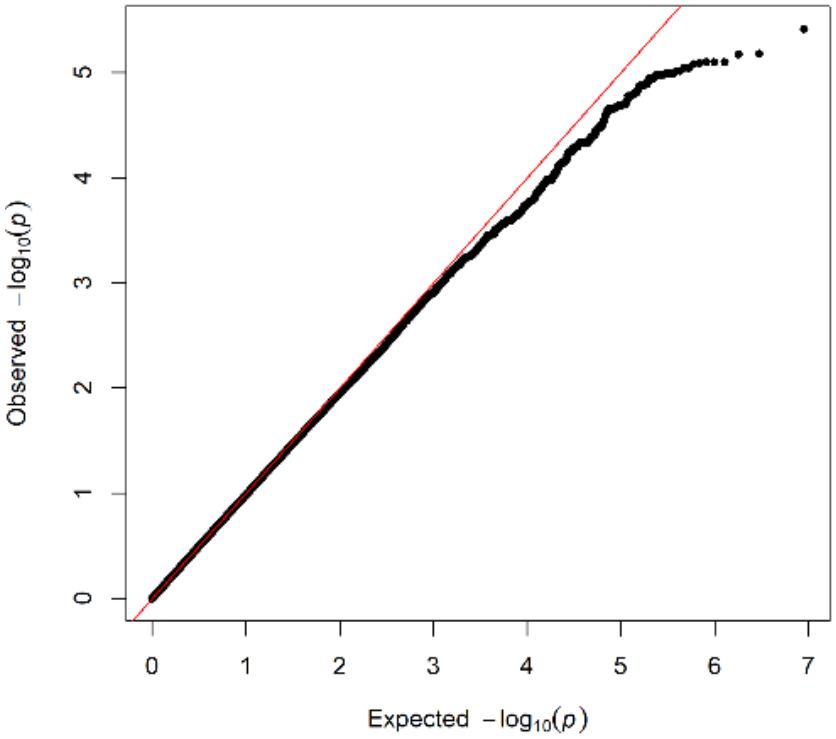
Supplementary Figure 8. Manhattan plot showing the SNPs with their p values that passed the 5×10^{-06} threshold (indicated by the red line) in the univariate Cox regression analysis (DSS: *additive* genetic model). Manhattan plot was generated using SNPs that satisfied the PH assumption in the univariate analysis. Note that rs7314075 is the SNP that passed the significance level of 5×10^{-08} in the multivariable analysis (see **Table 5.2**).



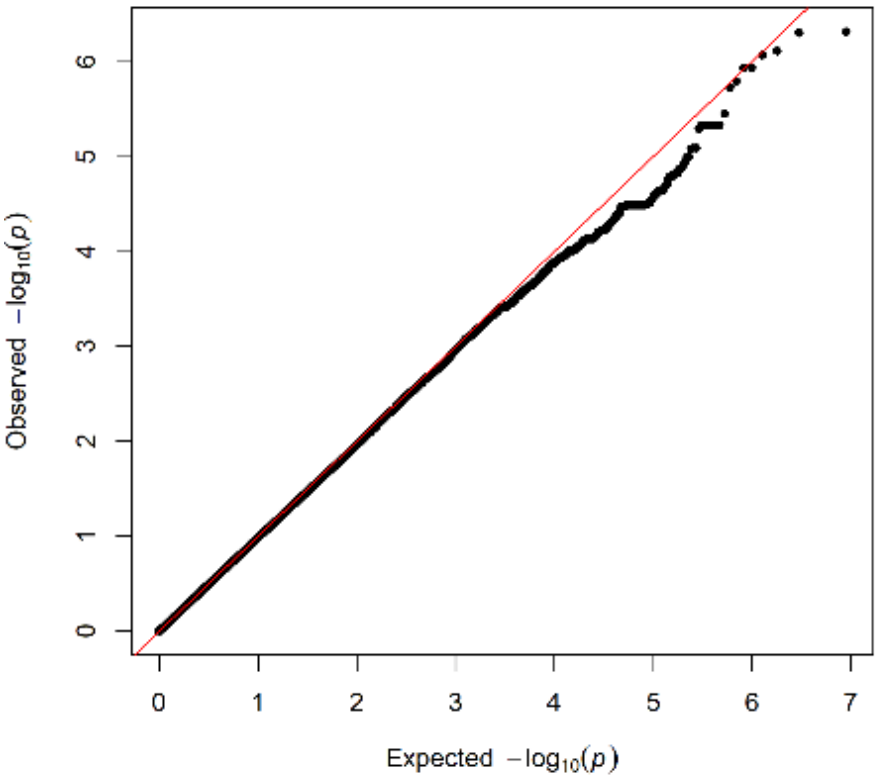
Supplementary Figure 9. Regional plot of rs7314075 in univariate analysis (DSS; *dominant* genetic model).



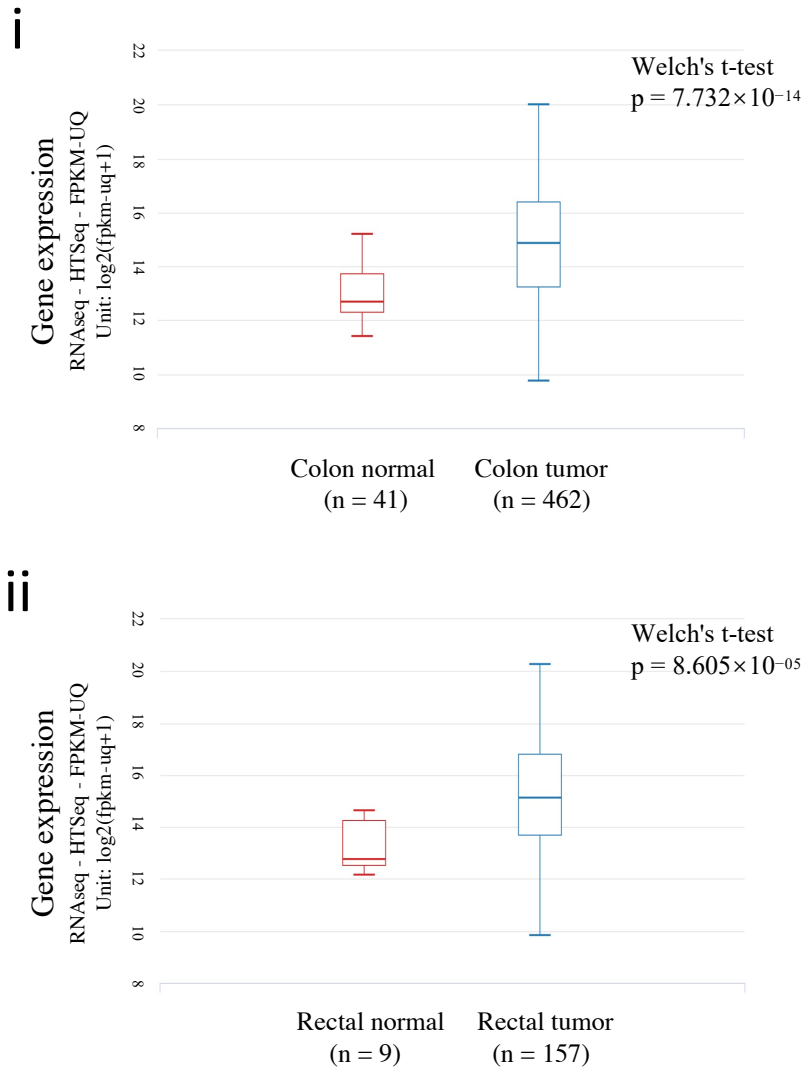
Supplementary Figure 10. Regional plot of rs7314075 in univariate analysis (DSS; *additive* genetic model).



Supplementary Figure 11. QQ plot for the univariate DSS analysis under the dominant genetic model. Plot was generated based on p-values of SNPs satisfied the PH assumption in the univariate analysis. Genomic inflation factor (λ) is 0.995.



Supplementary Figure 12. QQ plot for the univariate DSS analysis under the additive genetic model. Plot was generated based on p-values of SNPs satisfied the PH assumption in the univariate analysis. Genomic inflation factor (λ) is 0.987.



Supplementary Figure 13. Expression levels of *ERP27* in colorectal tumors and normal tissues. Analysis was done in UCSC Xena [72] using the GDC TCGA COAD and READ data. In both datasets, primary tumors and adjacent normal tissues (noted as "solid tissue normal" in TCGA data) were selected (recurrent and metastatic tumors were excluded), and only the tumors and normal tissues with their anatomical sites noted as colon (in COAD) and rectum and rectosigmoid junction (in READ) were analyzed. i, gene expression in colon tumors and normal tissues from TCGA COAD cohort; ii, gene expression in rectal tumors and normal tissues from TCGA READ cohort. Expression of *ERP27* is significantly higher in colon and rectal tumors than in normal tissues. The number of patients in the colon and

rectum tumor datasets is larger than those in the normal tissue datasets. This may explain why the gene expression levels in tumors have a higher variance compared to that in the normal tissues.

Appendix bibliography

1. Korpanty, G. J. *et al.* Association of BRM promoter polymorphisms and esophageal adenocarcinoma outcome. *Oncotarget* **8**, 28093–28100 (2017) doi:10.18632/oncotarget.15890.
2. Liu, G. *et al.* Two novel BRM insertion promoter sequence variants are associated with loss of BRM expression and lung cancer risk. *Oncogene* **30**, 3295–3304 (2011) doi: 10.1038/onc.2011.81.
3. Liu, G. *et al.* BRM Promoter Polymorphisms and Survival of Advanced Non-Small Cell Lung Cancer Patients in the Princess Margaret Cohort and CCTG BR.24 Trial. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **23**, 2460–2470 (2017) doi:10.1158/1078-0432.CCR-16-1640.
4. Segedi, M. *et al.* BRM polymorphisms, pancreatic cancer risk and survival. *Int. J. Cancer* **139**, 2474–2481 (2016) doi:10.1002/ijc.30369.
5. Wang, J. R. *et al.* Association of two BRM promoter polymorphisms with head and neck squamous cell carcinoma risk. *Carcinogenesis* **34**, 1012–1017 (2013) doi:10.1093/carcin/bgt008.
6. Pinto, D. *et al.* Functional impact of global rare copy number variation in autism spectrum disorders. *Nature* **466**, 368–372 (2010) doi:10.1038/nature09146.
7. Pinto, D. *et al.* Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat. Biotechnol.* **29**, 512–520 (2011) doi:10.1038/nbt.1852.
8. Winchester, L., Yau, C. & Ragoussis, J. Comparing CNV detection methods for SNP arrays. *Brief. Funct. Genomic. Proteomic.* **8**, 353–366 (2009) doi:10.1093/bfgp/elp017.

9. Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007) doi: 10.1101/gr.6861907.

10. Colella, S. *et al.* QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res.* **35**, 2013–2025 (2007) doi: 10.1093/nar/gkm076.

11. Diskin, S. J. *et al.* Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**, e126 (2008) doi:10.1093/nar/gkn556.

12. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002) doi:10.1093/nar/30.1.207.

13. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002) doi:10.1101/gr.229102.

14. Uddin, M., Sturge, M., Rahman, P. & Woods, M. O. Autosome-wide copy number variation association analysis for rheumatoid arthritis using the WTCCC high-density SNP genotype data. *J. Rheumatol.* **38**, 797–801 (2011) doi:10.3899/jrheum.100758.

15. Zheng, X. *et al.* Using family data as a verification standard to evaluate copy number variation calling strategies for genetic association studies. *Genet. Epidemiol.* **36**, 253–262 (2012) doi:10.1002/gepi.21618.

16. Lin, P. *et al.* Copy number variation accuracy in genome-wide association studies. *Hum. Hered.* **71**, 141–147 (2011) doi:10.1159/000324683.

17. Fernandez-Rozadilla, C. *et al.* A genome-wide association study on copy-number variation identifies a 11q11 loss as a candidate susceptibility variant for colorectal cancer. *Hum. Genet.* **133**, 525–534 (2014) doi: 10.1007/s00439-013-1390-4.

18. Marenne, G. *et al.* Assessment of copy number variation using the Illumina Infinium 1M SNP-array: a comparison of methodological approaches in the Spanish Bladder Cancer/EPICURO study. *Hum. Mutat.* **32**, 240–248 (2011) doi:10.1002/humu.21398.

19. Marenne, G. *et al.* Genome-wide CNV analysis replicates the association between GSTM1 deletion and bladder cancer: a support for using continuous measurement from SNP-array data. *BMC Genomics* **13**, 326 (2012) doi:10.1186/1471-2164-13-326.

20. Teo, S.-M. *et al.* A population-based study of copy number variants and regions of homozygosity in healthy Swedish individuals. *J. Hum. Genet.* **56**, 524–533 (2011) doi:10.1038/jhg.2011.52.

21. Tsuang, D. W. *et al.* The effect of algorithms on copy number variant detection. *PloS One* **5**, e14456 (2010) doi:10.1371/journal.pone.0014456.

22. Ukkola-Vuoti, L. *et al.* Genome-wide copy number variation analysis in extended families and unrelated individuals characterized for musical aptitude and creativity in music. *PloS One* **8**, e56356 (2013) doi:10.1371/journal.pone.0056356.

23. Jiang, L. *et al.* Genome-wide detection of copy number variations using high-density SNP genotyping platforms in Holsteins. *BMC Genomics* **14**, 131 (2013) doi:10.1186/1471-2164-14-131.

24. Nag, A. *et al.* CNV analysis in Tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1. *PloS One* **8**, e59061 (2013) doi:10.1371/journal.pone.0059061.

25. Degenhardt, F. *et al.* Association between copy number variants in 16p11.2 and major depressive disorder in a German case-control sample. *Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet.* **159B**, 263–273 (2012) doi:10.1002/ajmg.b.32034.

26. Need, A. C. *et al.* A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet.* **5**, e1000373 (2009) doi:10.1371/journal.pgen.1000373.

27. Priebe, L. *et al.* Copy number variants in German patients with schizophrenia. *PloS One* **8**, e64035 (2013) doi:10.1371/journal.pone.0064035.

28. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007) doi: 10.1086/519795.

29. Safran, M. *et al.* Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.* **31**, 142–146 (2003) doi:10.1093/nar/gkg050.

30. Haraksingh, R. R., Abyzov, A., Gerstein, M., Urban, A. E. & Snyder, M. Genome-wide mapping of copy number variation in humans: comparative analysis of high resolution array platforms. *PloS One* **6**, e27859 (2011) doi:10.1371/journal.pone.0027859.

31. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010) doi: 10.1038/nature09534.

32. Campbell, C. D. *et al.* Population-genetic properties of differentiated human copy-number polymorphisms. *Am. J. Hum. Genet.* **88**, 317–332 (2011) doi:10.1016/j.ajhg.2011.02.004.

33. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010) doi:10.1038/nature08516.

34. MacDonald, J. R., Ziman, R., Yuen, R. K. C., Feuk, L. & Scherer, S. W. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986-992 (2014) doi:10.1093/nar/gkt958.

35. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0. 2013-2015. Available at: http://www.repeatmasker.org (accessed on Nov 23, 2021).

36. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115 (2012) doi:10.1093/nar/gks596.

37. Arand, M. *et al.* A multiplex polymerase chain reaction protocol for the simultaneous analysis of the glutathione S-transferase GSTM1 and GSTT1 polymorphisms. *Anal. Biochem.* **236**, 184–186 (1996) doi:10.1006/abio.1996.0153.

38. Grambsch, P. M. & Therneau, T. M. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* **81**, 515–526 (1994) doi: 10.1093/biomet/81.3.515.

39. Quantin, C. *et al.* Variation over time of the effects of prognostic factors in a population-based study of colon cancer: comparison of statistical models. *Am. J. Epidemiol.* **150**, 1188–1200 (1999) doi: 10.1093/oxfordjournals.aje.a009945.

40. Green, R. *et al.* Very high incidence of familial colorectal cancer in Newfoundland: a comparison with Ontario and 13 other population-based studies. *Fam. Cancer* **6**, 53–62 (2007) doi: 10.1007/s10689-006-9104-x.

41. Woods, M. O. *et al.* The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut* **59**, 1369–1377 (2010) doi:10.1136/gut.2010.208462.

42. Yu, Y. *et al.* The long-term survival characteristics of a cohort of colorectal cancer patients and baseline variables associated with survival outcomes with or without time-varying effects. *BMC Med.* **17**, 150 (2019) doi:10.1186/s12916-019-1379-5.

43. Negandhi, A. A. *et al.* MTHFR Glu429Ala and ERCC5 His46His polymorphisms are associated with prognosis in colorectal cancer patients: analysis of two independent cohorts from Newfoundland. *PLoS One* **8**, e61469 (2013) doi: 10.1371/journal.pone.0061469.

44. Xu, W. *et al.* A genome wide association study on Newfoundland colorectal cancer patients' survival outcomes. *Biomark. Res.* **3**, 6 (2015) doi: 10.1186/s40364-015-0031-6.

45. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat. Methods* **9**, 179–181 (2012) doi:10.1038/nmeth.1785.

46. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009) doi:10.1371/journal.pgen.1000529.

47. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015) doi:10.1038/nature15393.

48. Crosslin, D. R. *et al.* Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. *Hum. Genet.* **131**, 639–652 (2012) doi:10.1007/s00439-011-1103-9.

49. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Front. Genet.* **5**, 370 (2014) doi:10.3389/fgene.2014.00370.

50. Kerminen, S. *et al.* Fine-scale genetic structure in Finland. *G3 Bethesda Md* **7**, 3459–3468 (2017) doi:10.1534/g3.117.300217.

51. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014) doi:10.1371/journal.pgen.1004234.

52. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004) doi:10.1038/nature03001.

53. Ahmad, M. *et al.* Inclusion of population-specific reference panel from India to the 1000 Genomes Phase 3 panel improves imputation accuracy. *Sci. Rep.* **7**, 6733 (2017) doi:10.1038/s41598-017-06905-6.

54. Manku, H. *et al.* Trans-ancestral studies fine map the SLE-susceptibility locus TNFSF4. *PLoS Genet.* **9**, e1003554 (2013) doi:10.1371/journal.pgen.1003554.

55. Namjou, B. *et al.* A GWAS study on liver function test using eMERGE network participants. *PLoS One* **10**, e0138677 (2015) doi:10.1371/journal.pone.0138677.

56. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat. Rev. Genet.* **11**, 499–511 (2010) doi:10.1038/nrg2796.

57. Werdyani, S. *et al.* Germline INDELs and CNVs in a cohort of colorectal cancer patients: their characteristics, associations with relapse-free survival time, and potential time-varying effects on the risk of relapse. *Cancer Med.* **6**, 1220–1232 (2017) doi: 10.1002/cam4.1074.

58. Kuningas, M. *et al.* Large common deletions associate with mortality at old age. *Hum. Mol. Genet.* **20**, 4290–4296 (2011) doi:10.1093/hmg/ddr340.

59. Kim, S.-Y., Kim, J.-H. & Chung, Y.-J. Effect of combining multiple CNV defining algorithms on the reliability of CNV calls from SNP genotyping data. *Genomics Inform.* **10**, 194–199 (2012) doi:10.5808/GI.2012.10.3.194.

60. Carter, N. P. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat. Genet.* **39**, S16-21 (2007) doi:10.1038/ng2028.

61. Tropeano, M. *et al.* Male-biased autosomal effect of 16p13.11 copy number variation in neurodevelopmental disorders. *PLoS One* **8**, e61365 (2013) doi:10.1371/journal.pone.0061365.

62. Campbell, M. K. & Ferrell, S. O. *Biochemistry*. (Thomson Brooks, 2009).

63. O'Brien, T. D., Jia, P., Caporaso, N. E., Landi, M. T. & Zhao, Z. Weak sharing of genetic association signals in three lung cancer subtypes: evidence at the SNP, gene, regulation, and pathway levels. *Genome Med.* **10**, 16 (2018) doi:10.1186/s13073-018-0522-9.

64. Wang, D. *et al.* Comparison of methods for correcting population stratification in a genome-wide association study of rheumatoid arthritis: principal-component analysis versus multidimensional scaling. *BMC Proc.* **3**, S109 (2009) doi:10.1186/1753-6561-3-s7-s109.

65. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015) doi:10.1186/s13742-015-0047-8.

66. Price, A. L. *et al.* Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83**, 132–135; author reply 135-139 (2008) doi:10.1016/j.ajhg.2008.06.005.

67. Polimanti, R. *et al.* A putative causal relationship between genetically determined female body shape and posttraumatic stress disorder. *Genome Med.* **9**, 99 (2017) doi:10.1186/s13073-017-0491-4.

68. Liu, J. *et al.* An integrated TCGA Pan-Cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018) doi:10.1016/j.cell.2018.02.052.

69. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930-934 (2012) doi:10.1093/nar/gkr917.

70. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012) doi:10.1101/gr.137323.112.

71. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013) doi:10.1038/ng.2653.

72. Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020) doi:10.1038/s41587-020-0546-8.