Variation in genes regulating angiogenesis, lymph-angiogenesis and metastasis: associations of three polymorphisms with outcome in patients with colorectal cancer

by

Lydia Annabel Dan

A thesis submitted to School of Graduate Studies in partial fulfillment of the requirement for the degree of

Master of Human Genetics

Discipline of Genetics/Faculty of Medicine

Memorial University of Newfoundland

October 2014

St. John's Newfoundland

Abstract

Biological and clinical findings show that the variation in the angiogenesis, lymph-angiogenesis and metastasis processes may affect patient survival. This study aims to identify new prognostic markers in colorectal cancer by investigating the associations of 381 genetic polymorphisms and haplotypes from 30 angiogenesis, lymph-angiogenesis and metastasis genes in a cohort of colorectal cancer patients from Newfoundland and Labrador. Our results showed that three linked SNPs located in the *MMP8* and *MMP27* genes were individually associated with overall survival (rs11225388, rs11225389, and rs12365082). By predicting and analyzing the haplotypes from these genes I also found an association between overall survival and an *MMP3* haplotype consisting of four polymorphisms. The biological consequences of these three SNPs and the *MMP3* haplotype and their relation to the risk of death in colorectal cancer are currently unknown. Future studies are required to replicate these findings in another cohort of colorectal cancer patients.

Acknowledgements

First and foremost my gratitude goes to my heavenly father Jehovah God, who gave me the strength that is beyond what is normal to push-on no matter the obstacles I face to complete this research project. I offer my sincere gratitude to my thesis supervisor, Dr. Sevtap Savas for her support, patient, motivation and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisor and mentor for my Master study. Thank you so much Dr. Savas for your guidance and mentorship. I would like to thank my thesis committee members: Dr. Roger Green and Dr. Jane Green for their encouragement, insightful comments and time.

I owe my deepest gratitude to the funding agents; the Canadian Institutes of Health Research (CIHR), Research and Development Corporation of Newfoundland (RDC), as well as the Research and Graduate Studies (RGS) for supporting me financially throughout the course of my duration in this program to the completion of my final thesis. I would like to thank Salem Werdyani for his help and support during the time of my research. I am indebted to many of my colleagues who supported me throughout the duration of my study. It is a pleasure to thank everybody who contribute in one way or another to the successful realization of my thesis, and as well as expressing my apology that I could not mention personally one by one.

And above all, my utmost thanks go to my family. Where would I be without my family? My parents deserved to be mentioned for their unflinching, inseparable support and prayers. Without their moral support and prayers I would have not gotten the strength to push-on. Thanks for being supportive and caring. Finally, I would like to thank my fiance, Simon Kuti for always cheering me up and standing by me through the good times and bad.

Table of contents

List of tables

List of figures

List of abbreviations

5-FU	5- Fluorouracil

A

AFAP	Attenuated familial adenomatous polyposis
AJCC	American joint committee on cancer
Ang	Angiopoietin
APC	Adenomatous polyposis coli

B

BMPR1A	Bone morphogenetic protein receptor, type IA

C

CAP	College of American Pathologists
CCND1	Cyclin D1
CEA	Carcinoembryonic antigen
CI	Confidence interval
CS	Cowden’s syndrome

D

dbCPCO	Database of colorectal cancer prognosis and clinical outcome
DCC	Deleted in colorectal cancer
DFS	Disease free survival
DNA	Deoxyribonucleic acid

E

ECM	Extracellular matrix
EPIC	European Prospective Investigation into Cancer and Nutrition

F

FAP	Familial adenomatous polyposis
FGF	Fibroblast growth factor
FGFR	Fibroblast growth factor receptor
FCCX	Familial colorectal cancer type X

H

HER-2/NEU	Human epidermal growth factor receptor 2
HIF	Hypoxia inducible factor
HNPCC	Hereditary non-polyposis colon cancer
HPPS	Hyperplastic polyposis syndrome
HRT	Hormone replacement therapy

HR	Hazard ratio
HWE	Hardy-Weinberg equilibrium

I

IGF	Insulin growth factor
IHC	Immunohistochemistry

J

JPS	Juvenile polyposis syndrome

K

KRAS	v-Ki-ras2 Kirsten rat sarcoma

L

LS	Lynch syndrome
LD	Linkage disequilibrium

M

MAF	Minor allele frequency
MAP	MUTYH-associated polyposis
MDS	Multidimensional scaling

MLH1	MutL homolog 1
MMR	Mismatch repair
MMP	Matrix metalloproteinase
MSI	Microsatellite instability
MTS	Muir-Torre syndrome
MSI-H	Microsatellite instability-high
MSI-L	Microsatellite instability-low
MSS	Microsatellite stable
MVA	Multivariable analysis

N

NFCCR	Newfoundland colorectal cancer registry
NSAID	Non-steroidal anti-inflammatory drug

O

OS	Overall survival

P

PAH	Polycyclic aromatic hydrocarbon
PCA	Principal component analysis
PCNA	Proliferating cell nuclear antigen

PDGF	Platelet derived growth factor
PDGFR	Platelet derived growth factor receptor
PFS	Progression free survival
PJS	Peutz Jeghers syndrome
PTEN	Phosphatase and tensin homolog
P53	Protein 53
P21	Protein 21
P27	Protein 27

Q

QC	Quality control

S

SMAD4	SMAD family member 4
SNP	Single nucleotide polymorphism
STK11	Serine/threonine kinase 11
SPSS	Statistical package for the social sciences

T

TGF	Transforming growth factor
TIMP	Tissue inhibitor of metalloproteinase

TNM	Primary tumor-T, Regional lymph node-N, Distant metastasis-M

U

UCSC	University of California, Santa Cruz
UICC	Union for International Cancer Control
UPA	Urokinase receptor

List of appendices

Research outputs and awards

Abstracts presented in conferences:

Oral presentation:

1) **Lydia A. Dan,** Jingxiong Xu, Salem Werdyani, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu, Sevtap Savas. Genetic polymorphisms in matrix metalloproteinase genes *MMP8* and *MMP27* are associated with overall survival in colorectal cancer. The TM's 3rd World Cancer Online Conference, January 21, 2014.

Poster presentations:

1) **Lydia A. Dan,** Jingxiong Xu, Salem Werdyani, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu, Sevtap Savas. Prognostic association of the polymorphisms in matrix metalloproteinase genes *MMP8* and *MMP27* in patients with colorectal cancer. The 3rd Annual Canadian Human and Statistical Genetics Meeting, May 3-6, 2014, Fairmont Empress, Victoria, British Columbia, Canada.

2) **Lydia A. Dan,** Jingxiong Xu, Salem Werdyani, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu, Sevtap Savas. Genetic polymorphisms in angiogenesis, lymph-angiogenesis, and metastasis pathway genes and the disease outcome in colorectal cancer. The 2013 Canadian Cancer Research Conference, November 2-6, 2013, Sheraton Centre Toronto, Ontario, Canada.

3) **Lydia A. Dan,** Jingxiong Xu, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Wei Xu, Roger Green, Sevtap Savas. Polymorphisms in vascular endothelial growth factor genes (*VEGFA*, *VEGFB* and *VEGFC*) and outcome in colorectal cancer. The 2nd Statistical and Human Genetics Conference, April 21-24, 2013, Esterel, Quebec, Canada.

Submitted abstract

1. Sevtap Savas, **Lydia A. Dan,** Jingxiong Xu, Salem Werdyani, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Geoffrey Liu, Roger Green, Wei Xu. Choosing the best genetic model to investigate the genetic polymorphisms in survival analyses. The 3rd International Conference on Translational Medicine, Las Vegas, USA, November 3-5, 2014 (Invited speech by Dr. Sevtap Savas).

Awards

Travel award

1) Travel award by the Beatrice Hunter Cancer Research Institute to attend the 2013 Canadian Cancer Research Conference ($996).

Graduate student award

2) The 2014 graduate student award for the Faculty of Medicine-Genetics based on the best publication record ($300) (Master`s category).

Chapter 1

1.1 Overview of the research study

Colorectal cancer is the fourth most common cause of cancer-related death worldwide accounting for over 9% of all cancers diagnosed (1, 2). The highest incidence of this disease in Canada is observed in Newfoundland and Labrador (NL) (3). The majority of colorectal cancers are sporadic while 5% to 10% of colorectal cancers are due to inherited mutations (4). Risk factors for colorectal cancer are classified as modifiable such as personal behavior and lifestyle factors or non-modifiable, such as age, family history, genetic factors and personal medical history (5).

Many factors that affect the outcome of colorectal cancer have been identified, but few of them have been robust or informative enough to provide guidance for clinical management. Prognostic factors that are well supported by research and that are used in patient management include tumor stage, age of diagnosis, regional node involvement, and residual tumor (6). In addition, many genetic markers have been identified as having prognostic or predictive utility for colorectal cancer outcome, but none have yet been integrated into patient management (6).

Recent studies have aimed to identify the genetic basis of prognostic variation in cancer patients. These studies usually investigate genetic variations, such as single nucleotide polymorphisms (SNPs), and their potential association with survival times. Among the candidate genes for such studies are the genes functioning in the angiogenesis,

lymph-angiogenesis and metastasis pathways. Angiogenesis is the growth of new blood vessels (7) while lymph-angiogenesis refers to the formation of new lymphatic vessels (8). Variations in these pathways may affect local tumor progression and distant metastasis (9, 10). Several genes, such as matrix metalloproteinases, have also been identified which may facilitate development of distant metastases in cancer patients.

The use of genetic polymorphisms such as SNPs and their combinations in haplotypes has helped to identify the genetic variations that can affect individual susceptibility to common complex diseases (11-13). Similar approaches, albeit in fewer studies, have been applied to identify genetic variations that are associated with the survival of cancer patients (14, 15).

The focus of this current research study is to analyze genetic polymorphisms and haplotypes in select candidate genes functioning in the angiogenesis, lymph-angiogenesis, and metastasis pathways in relation to survival outcomes of colorectal cancer patients. I tested the association of 381 polymorphisms and gene-based haplotypes with overall survival and disease-free survival in a cohort of 505 colorectal cancer patients from Newfoundland.

1.2. Introduction to colorectal cancer

Colorectal cancer is a disease caused by uncontrolled growth of cells within the colon or rectum (16). It is one of the most common gastrointestinal tract malignancies with a high incidence worldwide, especially in the developed countries (2, 17).

The development of colorectal cancer is a complex process involving multiple molecular pathways. Generally colorectal tumor growth is slow, yet tumors can spread to surrounding and distant tissues of the body (18). Since there may be no symptoms of colorectal cancer until late in the course of the disease, it is often diagnosed at an advanced stage (19). The prognosis in colorectal cancer varies with the extent of the disease at diagnosis: patients with early stage of the disease have longer survival times than those diagnosed with late stage and metastatic disease (20, 21). In recent years, mortality rates have fallen due to early detection, improved surgical techniques and adjuvant therapy (22).

1.3. Pathology of colorectal cancer

Colorectal cancer usually develops from normal mucosa to adenoma and then progresses to invasive carcinoma (23). Kinzler *et al.* (24) suggested that approximately 95 percent of colorectal tumors begin as a benign adenomatous polyp in the wall of the colon, developing into advanced adenomas, and then progressing to invasive cancer. The progression of this disease involves a series of genomic events, such as alterations in several oncogenes, tumor-suppressor and DNA repair genes, cell adhesion molecules, angiogenetic factors, and epigenetic factors (18). Colorectal tumors that are confined within the wall of the colon (stages I and II) are usually curable, but if left untreated they may spread to regional lymph nodes (stage III) or metastasize to distant sites (stage IV) (25).

1.4. Incidence and risk factors of colorectal cancer

Colorectal cancer accounts for over 9% of all cancers and as such is a significant cause of morbidity and mortality throughout the world (1, 2). Colorectal cancer is the third most common cancer worldwide and the fourth most common cause of cancer-related death (1). The incidence of colorectal cancer is not uniform throughout the world (2). Australia, New Zealand, Canada, United States, and parts of Europe are the countries with the highest incidence of this disease presumably because of the westernized diet and life-style. In contrast, developing countries such as China, India, and parts of Africa and South America have lower rates of this disease (2, 21). The incidence in developed countries is about 40 per 100,000 compared to five per 100,000 in Africa and some parts of Asia (1).

In Canada, colorectal cancer is the third most common cause of cancer-related death and the highest incidence of this disease is observed in NL (3). According to the report of the Canadian Cancer Society, it was estimated that approximately 23,900 Canadian would develop colorectal cancer in 2013 and 9,200 would die of this disease (12.7% of all cancer deaths) (3). The current estimated 5-year survival rate in Canada is 65% (3).

A number of factors contribute to the cause of colorectal cancer, including increasing age, nutritional factors, low physical activity, inflammatory bowel disease and genetic risk factors. Environmental risk factors are controllable, unlike hereditary factors and age. Evidence for the role of environmental risk factors comes from studies of those who migrate to other countries (26, 27). Those migrating from low-risk countries to high-risk countries have a tendency of having the increased risk of colorectal cancer typical of

the host population (26). For example, Japanese who migrate to Hawaii have increased risk of colorectal cancer compared with the Japanese who stay in Japan (26). One major reason for this is that western diets are high in fat, especially animal fat which is a major risk factor for colorectal cancer (2, 26). The EPIC (European Prospective Investigation into Cancer and Nutrition) study identified an increased risk of colorectal cancer in people with high consumption of meat (27). Other studies linked low folate (28-30) and low fiber consumption (31) with a higher risk of colorectal cancer. It is estimated that about 80% of all cases of colorectal cancer are caused primarily by diet. Thus changes in dietary habits might reduce the risk of this disease substantially (32).

Several other life-style factors are also associated with increased risk, including low levels of physical activity. Regular exercise increases metabolic rate and maximal oxygen uptake (33). Epidemiological studies show that men who are physically active have decreased risk of developing colorectal cancer (33). Cigarette smoking is another risk factor for colorectal cancer. Botteri *et al*. (34) reported that cigarette smoking is linked with formation and increased growth rate of adenomatous polyps which are precursor lesions of colorectal cancer. Regular consumption of alcohol may be associated with increased risk of colorectal cancer because reactive metabolites of alcohol, such as acetaldehyde, can be carcinogenic (35). Supporting this, another report suggested that those who are high consumers of alcohol also have diets low in essential nutrients, which can make their tissues more susceptible to carcinogenesis (1).

While dietary and other life-style factors may be controlled to some extent, colorectal cancer risk factors that an individual cannot control include age and hereditary factors. It is estimated that approximately 1% to 5% of colorectal cancer cases are linked

to highly penetrant genetic variants (36), such as the *APC* mutations in familial adenomatous polyposis (FAP), and mutations of DNA mismatch repair genes in Lynch syndrome (36, 37). In addition to these high-penetrance mutations, low penetrance alleles also contribute to the risk of colorectal cancer (38). The likelihood of developing colorectal cancer increases progressively from age 40 and rises sharply after age 50 (1, 21). More than 90% of colorectal cancer occurs in individuals aged 50 and over (21, 31).

1.5. Sporadic, hereditary and familial colorectal cancer

1.5.1 Sporadic colorectal cancer

Sporadic colorectal cancer development is multifactorial and is probably due to the combinations of numerous low-penetrant alleles and environmental or behavioral risk factors (39). In sporadic patients, there is no known familial history of colorectal cancer and age of diagnosis is usually late (median ~70 years) (40, 41). Low-penetrant alleles contribute modestly to the increase in colorectal cancer risk but when they interact with other susceptibility alleles or environmental factors they can modify the risk for colorectal cancer (42). Recently, several genome wide association studies (GWASs) have identified several single nucleotide polymorphisms (SNPs) that modestly influence the risk of colorectal cancer (43). Several meta-analyses have validated some of these genetic polymorphisms as susceptibility loci (38, 43, and 44).

1.5.2 Hereditary and familial colorectal cancer

A. Polyposis syndromes

Familial Adenomatous Polyposis (FAP)

FAP is an autosomal dominantly inherited syndrome caused by genetic mutations in the adenomatous polyposis coli gene (*APC*) (45, 46). It is characterized by the development of multiple (hundreds to thousands) adenomas in the rectum and colon after the first decade of life, resulting in colorectal tumors if not removed (46). Germline mutations in the tumor suppressor gene *APC* on chromosome 5q21 are the causes of FAP (47, 48). The APC protein is a part of a protein complex that targets β-catenin for degradation via GSK-3β–mediated phosphorylation (49). The median age of diagnosis of FAP is about 40 years, or 10 to 15 years after the initial development of polyposis (50, 51).

FAP exhibits close to 100% penetrance. More than 90% of patients with FAP will develop duodenal, ampullary, or peri-ampullary adenomas and 5% to 10% of the patients will develop duodenal carcinoma by the age of 60 (52, 53). A less aggressive but more variable variant of FAP is attenuated FAP (AFAP) characterized by fewer colorectal adenomatous polyps (usually 10 to 100) which is caused by mutations in the 3' part of *APC* (54). In some families with the mutations in 5’ end of the *APC* gene, the polyp burden is highly variable, from 10-20 polyps to 100s to 1000s polyps (55). Other variants of FAP are Gardner syndrome and Turcot syndrome. In Gardner syndrome numerous extracolonic features are observed, such as skin tumors, epidermoid cysts, congenital

hypertrophy of the retinal epithelium and desmoid tumors (56). This syndrome is also caused by mutations in the *APC* gene and may represent variable expression of a mutation also causing classic FAP (56). Turcot syndrome is a rare variant of FAP (57) in which patients develop polyposis and colorectal cancer along with central nervous system tumors (57). Studies associate Turcot syndrome with mutations in the DNA mismatch repair genes, *MLH1* and *MSH2* (57), and *APC* (58).

MUTYH-Associated Polyposis (MAP)

MUTYH-associated polyposis (MAP) is an autosomal recessive disorder characterized by adenomatous colon polyps and risk of colorectal cancer (59). It is caused by the mutation in the *MUTYH* gene (59). Patients with this disease typically develop 10–500 adenomas (59). MAP may account for 0.5% to 1% of all colorectal cancer cases (60). The age of onset of MAP has not been fully defined, but based on colorectal cancer cohort studies, it was suggested to be between ages 50 and 60 (61). A study by Jenkins *et al.* estimated that the lifetime risk for individuals with biallelic *MUTYH*-mutations to develop colorectal cancer is 80% (62). *MUTYH* is located on chromosome 1p34 (www.lovd.nl/MUTYH). It encodes a DNA glycosylase which plays a role in the DNA base-excision repair pathway (63). Two common *MUTYH* variants observed in MAP patients are the Tyr165Cys and Gly382Asp mutations (64).

Hyperplastic Polyposis Syndrome (HPPS)

HPPS is a rare condition that is characterized by the presence of multiple or large polyps throughout the colon (65). While it is inherited, no specific germ-line mutations or genetic abnormality have been noted in patients with HPPS (66). Individuals with this syndrome have a high risk of developing colorectal cancer (65). According to Young *et al.* (67) 50% of individuals with HPPS report a family history of colorectal cancer. Colorectal tumors in HPPS often have microsatellite stable (MSS) tumor phenotype (where mismatch DNA repair genes are not mutated) (66). Despite the different studies carried out, the mode of inheritance has not yet been completely determined, but based on the reports by Chow *et al.* (66) and Young *et al.* (67), either autosomal recessive or co-dominant is the most likely mode of inheritance.

B. Hamartomatous polyposis syndromes

Juvenile Polyposis Syndrome (JPS)

JPS is an inherited, autosomal dominant disorder distinguished by hamartomatous polyps in the gastrointestinal tract (68). Patients with JPS are likely to have various malignancies such as gastrointestinal, pancreatic, lung, uterine, ovarian and testicular tumors (69-71). About 68% of the JPS patients develop colorectal cancer by the age of 60 and average age of diagnosis of colorectal cancer is 42 (69). JPS is caused by germline mutations in the *SMAD4/DPC4* gene located on chromosome 18q21.1 and the *BMPR1A* gene located on chromosome 10q22-23 (72, 73).

Cowden's Syndrome (CS)

CS is another rare, autosomal dominant hamartomatous polyposis condition also characterized by tumors of breast, skin and thyroid (74). Germline mutations of *PTEN* are the cause of this disease (74). *PTEN* is a tumor suppressor gene and encodes a lipid phosphatase that regulates the PI3K/AKT pathway (75). Mutations in *PTEN* cause increased nuclear β-catenin that can lead to increased expression of c-Myc and cyclin D1 (*CCND1*) (75), two important cell signaling and cell cycle proteins with roles in carcinogenesis.

Peutz-Jeghers Syndrome (PJS)

PJS is an autosomal dominant syndrome leading to the development of gastrointestinal hamartomas and mucocutaneous hyper-pigmentation (76, 77). The overall incidence of colorectal carcinomas in PJS patients ranges from 20–50% (76, 77). Over their lifetime, patients with PJS have a 39% chance of developing colon cancer (76, 77). Germ-line mutations in *STK11 (LKB1)* are the cause of PJS. *STK11*, a tumor suppressor gene located on chromosome 19p13 (76, 77), encodes a serine-threonine kinase that modulates cell polarity and cell proliferation (76, 77).

C. Hereditary non-polyposis colon cancer

Hereditary non-polyposis colorectal cancer (HNPCC) can be sub-divided into two categories: Lynch syndrome (LS), which is caused by the DNA mismatch repair (MMR) gene mutations; and familial colorectal cancer type X (FCCX). The genetic causes of

FCCX is currently unknown (78), but likely there are many different genes mutated in different families.

LS is an autosomal dominant condition that is responsible for 2% to 5% of all colorectal carcinoma cases (78, 79). Lynch syndrome is caused by germline mutations in one of the several MMR genes such as *MSH2*, *MLH1*, *MSH6* and *PMS2* (80-88). These MMR genes encode proteins that help maintain the integrity of short segments of nucleotide repeats known as microsatellite sequences (80-88). When MMR genes are mutated, the encoded proteins are unable to repair bases that are incorrectly added to or deleted from microsatellite sequences during DNA replication (88). Thus, colorectal tumours in LS patients are characterized by microsatellite instability (MSI) (89). MMR mutation carries have a 50–80% lifetime risk of developing colorectal cancer, 50–60% risk of developing endometrium carcinoma (in women), and up to 15% risk of other tumors such as tumors of stomach, ovary, hepatobiliary tract, upper urinary tract, pancreas, small bowel and central nervous system (78). Abdel-Rahman *et al*. (78) reported that the median age of colorectal cancer diagnosis in Lynch syndrome patients is 44.

FCCX patients meet the Amsterdam criteria I (briefly, early age of diagnosis and multiple individuals affected in more than one generation) but show no evidence of MMR gene defect (90). Patients with FCCX have increased risk of colon cancer, but usually not of the other cancers that are typical of Lynch syndrome (90). The average age of onset is about 60 years, which is higher than in LS (90). In spite of intensive research, the genes for FCCX have so far remained unidentified (78). It is also possible that some or many cases of FCCX are due to clustering of sporadic colorectal cancer.

1.6. Prognostic markers in colorectal cancer

According to the definition by the National Cancer Institute, "prognosis is an estimate of the likely course and outcome of a disease" (31). There are increasing numbers of prognostic factors that have been identified over the years, some of which may be used in outcome predictions and management decisions. Prognostic factors that have been repeatedly investigated include stage, age at diagnosis, residual disease, histologic type and grade, carcino-embryonic antigen (CEA) levels, extramural venous invasion, and submucosal vascular invasion in malignant polyps (6). Many molecular, protein, and carbohydrate markers have been investigated as possible prognostic factors, but so far none has been integrated into patient care (6).

In 1999, the College of American Pathologists (CAP) evaluated the prognostic roles of pathologic, genetic, molecular, and other biological factors in colorectal cancer (6). Putative prognostic factors were grouped into categories that reflected the strength of the published evidence demonstrating their prognostic value (6).

1.6.1 Category I: prognostic markers used for management of colorectal cancer patients

Category I markers were defined by CAP as the best indicators of prognosis for colorectal cancer and include tumor stage, regional node involvement, vascular invasion, and residual tumor (6). This group of prognostic factors are those that are well documented with evidence from multiple published and statistically robust trials and are used clinically (6).

Tumor stage (defined based on the tumor characteristics) and disease stage (defined based on both the tumor characteristics and the presence or absence of metastases detected by diagnostic imaging) are well-established prognostic markers used in the clinic; they indicate the extent of the disease (i.e. size of the tumor, the depth of tumor penetration or metastatic disease) and influences survival outcomes of patients (91-93). Survival in colorectal cancer is highly dependent upon the stage of the disease at diagnosis. The 5-year survival rates are about 90% for stage I (early stage), 70% for regional tumors (stage II and III) and 10% for people diagnosed with distant metastatic cancer (stage IV) (94). Accurate staging is very critical for appropriate patient management and meaningful clinical research (95). Although a large number of staging systems have been developed for colorectal cancer over the years, only the TNM (Primary tumor-T, Regional lymph node-N, and Distant metastasis-M) staging system of American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (IUAC) is widely recommended (96, 97).

Regional node involvement, which is a part of the TNM staging, is a strong predictor of outcome in colorectal cancer (6). TNM classifies nodal involvement as a prognostic marker in colorectal cancer based on the number of cancer-invaded lymph nodes (96, 97). Reports show that the number of lymph nodes obtained during surgery is critical for the prognosis of stage II and stage III colon cancer patients (98, 99) as it helps with accurate TNM staging. Expert groups recommend at least 12 nodes be examined histologically to accurately determine the nodal status (6, 100).

Another important prognostic determinant in colorectal cancer is the lymphatic or vascular invasion. In these cases, tumor invasion occurs in veins or in small non-

muscularized vessels that represent either post-capillary lymphatics or venules (6). Invasion of tumor cells into lymph or blood vessels is a (crucial) step in the metastatic process (101). Lymph node metastases and distant metastases are common in advanced colorectal cancers (102-105). Several studies suggest that venous invasion and lymphatic invasion may be independent prognostic factors in colorectal cancer (106-108).

The amount of residual tumor is a prognostic factor (6). IUIC (109) and AJCC (110) classify residual tumor (R) as: R0, no residual tumor; R1, microscopic residual tumor; and R2: macroscopic residual tumor. The better the original tumor is removed during the surgery (e.g. R0), the lower the recurrence risk.

In essence, the recommendations made by CAP regarding the prognostic factors are the best opinion. However, despite the enormous number of studies exploring the prognostic significance of various histologic, molecular, and clinical features, clinical stage at diagnosis remains the best indicator of the prognosis for colorectal cancer.

1.6.2 Category IIA and IIB: prognostic markers with good evidence but not in use for clinical management of colorectal cancer patients

Based on the CAP guidelines on prognostic factors in colorectal cancer, category IIA markers are potential prognostic markers with good evidence but their importance for clinical use is not yet established. Such markers include histologic tumor type, tumor grade, and MSI status (6).

Based on previous studies, the signet-ring cell type of adenocarcinoma and small-cell carcinoma are the only histologic types of colonic carcinoma that consistently have

been found to have stage-independent adverse effects on prognosis (111). However, usually the establishment of prognostic value of histologic type is hampered by the insufficient amount of data extracted during the pathological examination of tumor tissues (6).

Tumor grade is another prognostic marker with strong evidence but not in use in the clinic (6). Tumor grade is the degree of tumor differentiation and in some studies has been demonstrated to be a stage-independent prognostic factor in colorectal cancer (112). In the majority of the studies, the prognostic significance of grade is investigated in statistical analysis as low grade (well and moderately differentiated) versus high grade (poorly differentiated or undifferentiated) (6). CAP and AJCC/UICC recommended the adoption of this two-tiered grading system for colorectal cancer (6, 96, and 97). However, despite the number of grading systems that have been suggested in the literature, there is no single widely accepted and employed standard for tumor grading (113, 114).

Last but not least, MSI is considered as a category II prognostic marker for colorectal cancer (6). There are three types of MSI tumor phenotype; MSI-H (MSI-high), MSI-L (MSI-low), and MSS (microsatellite-stable). Studies show that patients with MSI-H tumor phenotype have better prognosis when compared to patients with MSS and MSI-L tumor phenotypes (115).

1.6.3 Category III: genetic markers as potential prognostic markers in colorectal cancer

Many genetic and molecular markers have been identified as having potential prognostic or predictive utility for colorectal cancer (6). These potential markers are those

listed by CAP under the category III include molecular markers, markers of cell proliferation or angiogenesis, and proteases (6). Large prospective cooperative group studies are currently ongoing that will clarify the prognostic value of many of these factors (6). **Table 1.1** shows some of the potential prognostic and predictive genetic markers studied in colorectal cancer (6). Below, some of the well-studied markers are discussed in detail.

KRAS

KRAS is a member of the *RAS* oncogene family (116-118). Mutation of *KRAS* occurs in approximately 50% of colorectal tumors (119). *KRAS* mutation occurs during adenoma progression, after *APC* mutation (120). Some *KRAS* mutations are predictive of a worse outcome and are associated with recurrence of colorectal cancer after therapy (121). However, other studies have failed to demonstrate any statistically significant link between *KRAS* mutations and prognosis (122). Several large studies have also failed to demonstrate the effect of *KRAS* mutations on disease-free or overall survivals, either in isolation or in combination with other mutations (123).

TP53

The *TP53* gene is located on the short arm of chromosome 17 (17p13.1) (124). The function of *TP*53 includes control of the cell cycle, DNA repair and synthesis, genomic plasticity and programmed cell death (124). That is why it is called the ‘guardian of the genome’ (125).

Table 1.1: Potential prognostic and predictive genetic markers in colorectal cancer (6)

Candidate Biomarkers
KRAS
TP53
DCC/18q
NM23
APC
SMAD4
BRAF
MLH1
TYMS
TIMP
VEGF
CD44
Matrix metalloproteinases (MMPs)
BCL-2
BAX
TYMP
MSI
CEA levels
C-reactive protein levels

TP53 mutations are the most common genetic alterations reported in human cancers (126). In colorectal adenomas, *TP53* mutations or allelic loss occur as late events in tumor progression (127). There are studies suggesting the prognostic and predictive significance of *TP53* mutations in colorectal cancer. For example, Tortola *et al*. (128) showed that mutations in *TP53* were predictive of worse outcome. Yamaguchi *et al*. (129) concluded that patients with *TP53* mutated tumors had a five-fold higher recurrence rate

and risk of death. However, despite these results, many other studies have failed to identify the prognostic effect of *TP53* in colorectal cancer. For example, Soong and coworkers studied 995 patients with Dukes' B and C colorectal cancer tumors, and no prognostic significance of the *TP53* mutations was observed (130). Similarly, the study reported by Elsaleh *et al.* (131) failed to identify an effect of *TP53* mutations on prognosis or therapeutic response to adjuvant chemotherapy in patients with Dukes' C tumors. Therefore, currently there is no convincing evidence of the prognostic role of *TP53* mutations in colorectal cancer.

DCC/18q

DCC (deleted in colorectal cancer) is a gene located on the long arm of chromosome 18 (18q) (132). Cytogenetic studies demonstrated that deletions of chromosome 18q were relatively common in colorectal cancer (133). In some studies, *DCC*/18q deletion was suggested as a useful prognostic marker (134). However, other studies using similar techniques have failed to confirm the prognostic association of loss of *DCC* in patients with colorectal cancer (135).

NM23

NM23 genes are located on chromosome 17 (17q21.3) and two of these genes are found in humans, namely *NM23-H1* and *NM23-H2* (136). *NM23* genes are putative metastatic suppressor genes (136). In advanced cases of colorectal carcinoma, somatic deletions of the *NM23* genes have been reported. Campo *et al*. (137) identified the deletions of *NM23-H1* in 56 patients with aggressive behavior of colorectal carcinomas.

Similar findings have also been reported by others, showing that over-expression of *NM23-H1* is significantly reduced in patients with advanced disease compared with patients with earlier disease stages (138). However, many other studies have failed to demonstrate a prognostic role of *NM23-H1* expression in colorectal cancer (139).

1.7. Angiogenesis, lymph-angiogenesis and metastasis

Angiogenesis is the formation of new blood vessels from an existing blood vessel (7). Events included in this process are proliferation, migration, and invasion of endothelial cells, organization of endothelial cells into functional tubular structures, maturation of vessels, and vessel regression (7). Tumor cells cannot grow beyond a critical size or metastasize to another organ without the formation of new blood vessels around the cells (7).

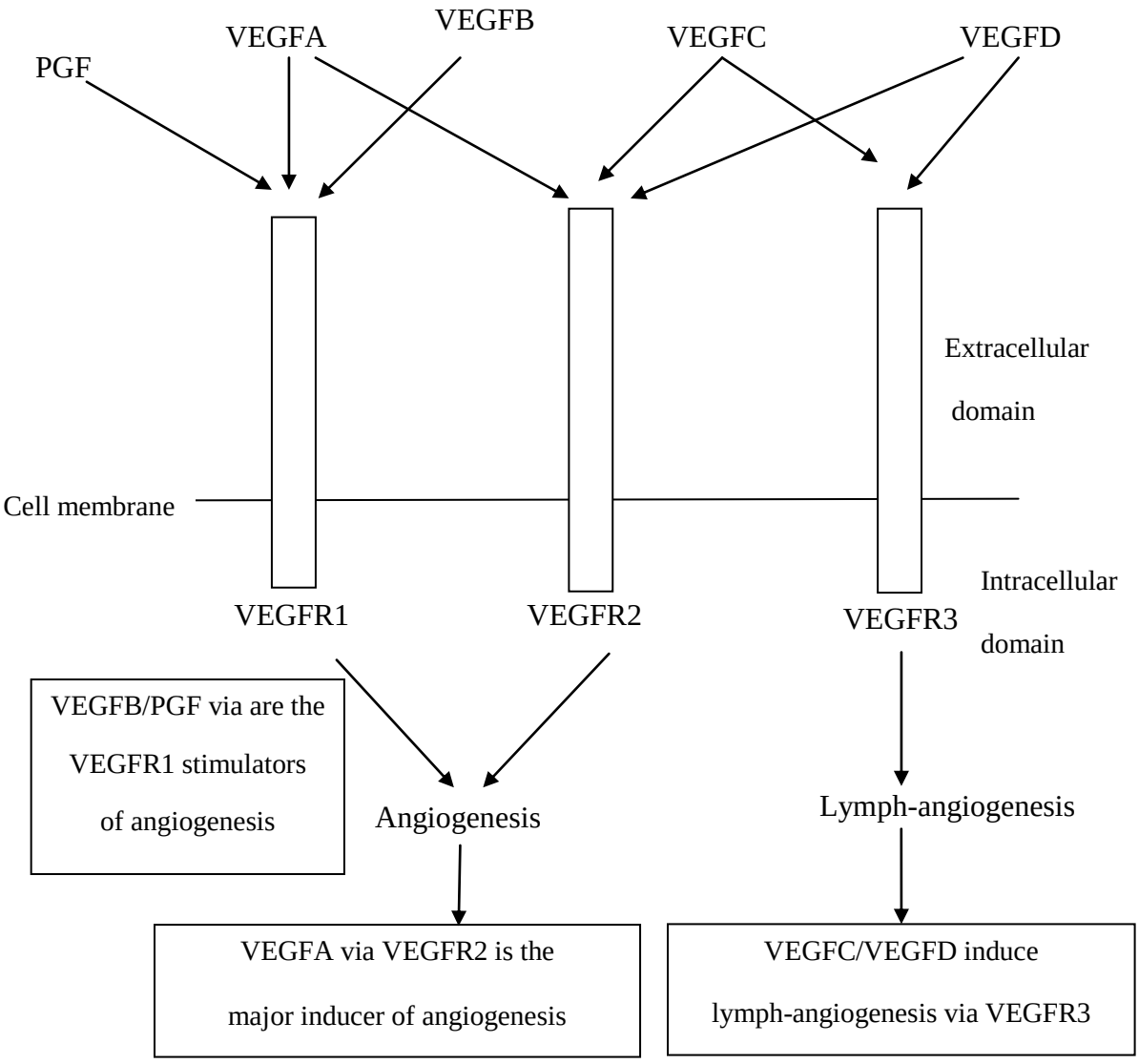
Angiogenesis around tumors was observed many years ago (7, 140). In 1971, Folkman proposed that tumor growth and metastasis are angiogenesis-dependent, and that if angiogenesis is blocked, then that could help arrest tumor growth (7). Since then, intensive search has been done for pro- and anti-angiogenic molecules. Research published by Gullino in 1976 showed that cells in pre-cancerous tissue acquire angiogenic capacity on their way to becoming cancerous (141). It is now a widely accepted concept that angiogenesis is "on" when pro-angiogenic molecules are activated and is "off" when they are inhibited (142). Signals that trigger this switch have been discovered by research involving hypoglycaemia, mechanical stress generated by proliferating cells, immune/inflammatory response (i.e. immune/inflammatory cells that infiltrate the tumor)

and genetic mutations that lead to the activation of oncogenes or inactivation of tumour-suppressor genes that control production of angiogenesis regulators (142, 143). Angiogenesis is regulated by many growth factors such as vascular endothelial growth factors (VEGFs), platelet-derived growth factor (PDGF), insulin-like growth factor (IGF), transforming growth factor (TGF), angiopoietins (Angs), and several chemokines (144, 145). Among these, VEGFs have a predominant role as the key regulators of angiogenesis (146). The interaction of the VEGFs and placental growth factor (PGF) family members with cell surface receptors (VEGFRs) leads to cascades of signaling that lead to the formation of new blood and lymphatic vessels (147) (**Figure 1. 1**).

Lymph-angiogenesis is the growth of new lymphatic vessels from an existing lymphatic vessel (8). Lymph nodes play an essential role in both normal and pathologic conditions (10, 148). In brief, under normal conditions the main functions of the lymph nodes are to remove excess fluid from the blood circulation, to transport immune cells that help trap infectious agents, and in cancer, to carry cancer cells to the lymphoid tissues and beyond (10). However, in cancer various studies show that most human tumors are able to metastasize via the lymphatic or blood vessels to other tissues in the body (149, 150).

Expression of lymph-angiogenesis-inducing growth factors in a range of animal tumor models has been well studied (151, 152). The signaling system consisting of VEGFC and VEGFD binding to VEGFR3 is a well-known mechanism of action behind lymph-angiogenesis (148) (**Figure 1. 1**).

Figure 1.1: Interactions between the VEGF ligands and their receptors (VEGFRs) (147, 148, and 153).



VEGF receptors (VEGFR1, VEGFR2, and VEGFR3) are shown as vertical rectangles. Copyright permission by the publisher, Elsevier, of the journal "Current Opinion in Cell Biology".

Vascular Endothelial Growth Factors (VEGFs) and their receptors (VEGFRs)

VEGF ligands and their receptors bind together to activate cellular signals for angiogenesis and lymph-angiogenesis. VEGF ligands and receptors are the most intensely investigated proteins in cancer as they play crucial roles in both normal and pathologic angiogenesis and lymph-angiogenesis (153, 154). The VEGF family of ligands are VEGFA, VEGFB, VEGFC, VEGFD and PGF and the three VEGF receptors are VEGFR1, VEGFR2 and VEGFR3 (**Figure 1.1**) (153). VEGFs are up-regulated by hypoxia-inducible factor 1α (*HIF1α*), and extracellular matrix (ECM) for the purpose of initiating an angiogenic switch that promotes tumor growth (153).

Among the VEGF ligands, VEGFA is the most well characterized one (153). The mechanism behind the biological effect of VEGFA involves its interaction with the cell surface receptors VEGFR1 and VEGFR2 located on the vascular endothelium (153) (**Figure 1.1**). Their interactions play a crucial role in angiogenesis, which is critical for cancer progression (153). For example, in breast cancer, increased production of VEGFA is correlated with early relapse (155).

VEGFB is one of the least characterized members of the VEGF family of ligands. It was discovered a few years after VEGFA and PGF (156). VEGFB exists in two isoforms, which bind to VEGFR1 but not to VEGFR2 or VEGFR3 (**Figure 1.1**) (157). It is expressed in the endothelial and mural cells, skeletal muscle, adipose tissue, and smooth muscle cells in adults (156, 158). According to studies, VEGFB is detectable in many tumors including colorectal, meningioma, lung and breast tumors (159-161).

VEGFC is another VEGF ligand. VEGFC binds to VEGFR2 and VEGFR3 (**Figure 1.1**) (162). The VEGFC signaling via VEGFR2 and VEGFR3 plays a critical role

in cancer progression (163). Mandriota *et al.* (164) showed that VEGFC is involved in tumor lymph-angiogenesis through inducing the formation of additional lymphatic vessels by which tumors cells find a channel to metastasize to distant sites. Further reports show that VEGFC is involved in the progression of several types of malignant tumors such as lung, colorectal, and breast tumors (165-167).

VEGFD is another ligand of the VEGF family. It stimulates the growth of vascular and lymphatic endothelial cells by signaling via VEGFR2 and VEGFR3 (168) (**Figure 1.1**). VEGFD is expressed in the adult lung, heart, muscle, and small intestine, but mostly found in the foetal lungs and skin (168). Expression of VEGFD in many tumor types has been detected, and it has been implicated to have a role in tumor angiogenesis and lymph-angiogenesis in breast cancer (169), esophageal squamous cell carcinoma (170), and lung cancer (171). Expression of VEGFD has also been implicated as a poor prognostic marker for colorectal (172), ovarian (173), gastric (174), and lung cancers (175).

PGF, placental growth factor, is another VEGF ligand (176). It is expressed in the placenta, heart and lungs (177). So far, four human PGF isoforms have been reported (178). PGF binds to the cell surface receptor VEGFR1 located on the vascular endothelium, which can stimulate angiogenesis (179) (**Figure 1.1**). It helps in the growth, migration and survival of the endothelial cells (178, 179). A report by Fischer and coworkers showed that PGF is involved in various pathological conditions such as tumor growth, arthritis, ocular ischaemia, and obesity (180). Wei *et al*. (181) linked PGF expression with disease progression in colorectal cancer. The Chen *et al*. (182) report correlates tumor stage and patient survival in gastric cancer. Also, elevated levels of PGF

expression were associated with recurrence, metastasis and patient mortality in breast cancer (183).

VEGFR1, also known as FLT1, is a cell surface receptor expressed at high levels in the vascular endothelial cells throughout fetal development and in the adult tissues (180, 184). It is activated when VEGFA, VEGFB or PGF binds to it (180, 184) (**Figure 1.1**). VEGFR1 helps the migration of the endothelial cells (180). VEGFR1 has been found to be expressed in various types of malignant cells such as colorectal, prostate, breast, esophageal cancers and leukemia (185-188).

VEGFR2, also known as KDR in humans, is a cell surface receptor that plays a very important role in the development of endothelial cells (189). It is expressed in the vascular endothelial cells located on the vascular endothelium (189). VEGFR2 can be activated when VEGFC or VEGFD ligands bind to it (189) (**Figure 1.1**). Shibuya describes VEGFR2 as a major inducer of angiogenesis as it helps promote endothelial cell differentiation, proliferation, migration and formation of new vascular vessels (190). VEGFR2 is implicated as a prognostic marker in patients with different types of malignancies including endometrial carcinoma and colorectal cancer (191, 192).

VEGFR3 is also a cell surface receptor located on the vascular endothelium. It is coded by the *FLT4* gene in humans. VEGR3 is activated when VEGFC or VEGFD ligands bind to it (162, 168) (**Figure 1.1**). Its major function is to induce lymphatic endothelial cells to form new lymphatic vessels (162, 168). According to literature findings, the interaction of VEGFC and VEGFR3 plays a role in disease progression and lymph node metastasis in prostate cancer (193). Other studies have also reported

VEGFR3 as involved in the progression of several types of malignant tumors, such as colorectal, breast, and melanoma tumors (192, 194 and 195).

Metastasis is the spread of cancer cells from the primary tumor site to the lymph nodes or to other tissues in the body (e.g. liver, brain). Abnormalities in tumor angiogenesis and lymph-angiogenesis are the key causes of this often deadly problem (10, 140 and 147). Studies suggested that the spread of primary tumor cells to distant organs depends critically on the formation of new blood vessels and lymphatic vessels, because these vessels not only provide oxygen and nutrients, remove waste materials from the tumor but, also provide a route of exit for tumor cells into the blood stream or lymph nodes (145, 147).

The most convincing correlation between angiogenesis and tumor metastasis has been reported in cases where vascular density of tumors has been correlated with metastasis and patient outcome. Weidner *et al.* (196) showed a direct correlation between the vascular density and the risk of metastasis in breast cancer patients. Other groups have repeated this study and most have confirmed the initial correlation not only in breast cancer but also in tumors of other tissues such as prostate, lung, stomach, and cervix (196-200).

For tumor cells to metastasize, they must detach themselves from the tumor. The degradation of the extracellular matrix (ECM) plays a critical role in this process (201). Many reports show that one of the hallmarks of cancer cells is the alteration of their interactions with the ECM, which is induced either by the tumor cells or by surrounding cells such as fibroblasts, macrophages and leukocytes (201). The ECM can regulate tumor cell growth by binding to and storing cytokines, by promoting cell attachment and

migration, by providing a stable foundation, supporting cell growth and survival by interacting with cell-surface receptors, and by activating appropriate signaling pathways (202, 203). According to Chambers and Matrisian, matrix metalloproteinases (MMPs) are implicated in the progression of many human cancer types because they help the degradation of the ECM, thus helping cancer cells to spread to distant organs which are the main cause of death in patients with malignant disease (204).

According to the HUGO database (205), there are 23 MMP genes in the human genome. Several studies investigated the roles of MMPs in cancer progression. For example, one study showed that high serum levels of *MMP9* was associated with rapid progression, poor survival and secondary metastasis in patients with melanoma (206). In other studies, lymph node metastases and poor outcome was associated with the tumor levels of *MMP9* and *MMP2* in patients with laryngeal cancer (207, 208). In summary, these and previously discussed literature findings suggest that in addition to VEGFs and VEGFRs, MMPs may also play crucial roles in cancer progression.

1.8. Angiogenesis, lymph-angiogenesis and metastasis pathways and prognosis in colorectal cancer

The connection between the genes functioning in angiogenesis, lymph-angiogenesis, and metastasis processes and prognosis in colorectal cancer has been investigated intensively over the years. The majority of these studies focused on VEGFA. For example, high levels of *VEGFA* expression in metastatic human colon carcinomas have been reported to correlate with poor prognosis in patients (209). In another report,

VEGFA expression was found to be higher in metastatic tumors than in non-metastatic tumors, and was correlated with liver metastasis and poor patient prognosis (210). Takahashi *et al.* (211) showed that colon cancer patients with tumors with increased *VEGFA* levels have significantly shorter 5-year disease-free survival (DFS) times. Cascinu *et al.* (212) confirmed this finding. Another study reported the relation of high *VEGFA* expression with progression in colorectal cancer where a greater intensity of VEGFA staining was associated with greater lymph node metastasis, higher stage, and shorter disease-specific survival; based on these results the authors concluded that *VEGFA* expression in colorectal cancer appears to be an independent prognostic marker of tumor behavior and can be useful in identifying patients with unfavourable clinical outcome (213).

Other studies reported the prognostic significance of the serum VEGFA levels in colorectal cancer. An example of this was a large study conducted by the Danish Colorectal Cancer Study Group (214) where high preoperative VEGFA concentrations were associated with reduced overall survival times in patients with colon carcinoma (214). In addition, De Vita *et al.* (215) reported that preoperative serum VEGFA level might be useful for predicting outcome in patients with colon cancer who undergo surgery.

Although not intensely studied, other VEGF family ligands have also been reported to be associated with the progression of colorectal cancer. In one study, PGF levels were reported to be associated with disease progression and patient survival in colorectal cancer (216). Jayasinghe *et al.* (217) reported that VEGFB promotes tumor survival and thus helps progression of colorectal cancer while White *et al.* (218) reported

that the expression of VEGFD was associated with lymphatic involvement and reduced patient survival in colorectal carcinoma. Also, Rmali *et al*. (219) reported a correlation of VEGFR2 expression with disease progression in colorectal cancer patients.

The matrix metalloproteinases (MMPs) have also been implicated in the progression of colorectal cancer. Several studies reported over-expression of *MMP1, MMP2, MMP3, MMP7, MMP9,* and *MMP13* in colorectal tumors (220). One report showed that high levels of *MMP3* expression in colorectal cancer were associated with poor prognosis (221). Further, a meta-analysis highlights the prognostic effect of *MMP9* in colorectal cancer patients; in this analysis patients with higher tumor expression of *MMP9* were found to have poorer survival (222). Another study, including a meta-analysis suggested that tumor *MMP2* expression is an independent prognostic factor in colorectal cancer patients (223; 224). Yang *et al.* (225) reported that over-expression of *MMP12* can predict outcome in patients with colorectal cancer. These and other literature findings suggest a critical role of VEGFs, VEGFRs and MMPs in prognosis and progression of colorectal cancer.

1.9. Genetic polymorphisms in angiogenesis, lymph-angiogenesis and metastasis pathway genes and their relation to progression in colorectal cancer

A number of studies analyzed genetic polymorphisms in VEGF ligand and receptor genes and MMP genes in relation to the prognosis of colorectal cancer patients. The majority of these studies are summarized in the public dbCPCO database (database of colorectal cancer prognosis and clinical outcome) (226).

According to the dbCPCO database, a number of *VEGFA* polymorphisms have been examined in different studies, but often reported conflicting results. As an example, Dassoulas *et al*. (227) reported that one *VEGFA* SNP (-634G/C; NM_001025366.1:c.-94C>G) was associated with overall survival (OS) in colorectal cancer patients. However, many other studies did not find this association in their cohorts (228-232). Similarly, Zhang *et al.* (233) showed no association of another *VEGFA* SNP (+936C/T; NM_001025366.1:c.*237C>T) with OS or DFS, yet Dassoulas *et al.* (227) reported an association of this polymorphism with prognosis. For another *VEGFA* SNP (-1498C T/C in promoter; NG_008732.1:g.4534C>T), associations with OS and DFS in stage II patients and progression free survival (PFS) and OS in metastatic colorectal patients was reported (234), however other groups did not replicate these findings (232, 235). In the case of the -2578C/A polymorphism (NG_008732.1:g.3437A>C), no association was detected with OS and DFS (231) or with PFS in colorectal cancer patients (232). As of October 2013, there were no entries in the dbCPCO database regarding polymorphisms in other VEGF ligand genes and prognosis in colorectal cancer.

Among the VEGFRs, KDR is frequently studied in prognostic studies in colorectal cancer. Hansen *et al*. (236) investigated the prognostic effect of a *KDR* polymorphism (-604 T/C; NM_002253.2:c.-906T>C) and reported its association with PFS in a cohort of colorectal cancer patients, but conflicting results regarding this polymorphism were also reported (232, 237 and 237). In addition, association of another *KDR* SNP (1719 A/T; NP_002244.1:p.Gln472His), with survival was identified in multiple studies (236, 237). Lastly, one study that analyzed the prognostic effect of the *VEGFR1*-519C/T genetic variation did not find association with patient survival (238).

According to the dbCPCO database, a small number of studies were conducted investigating the polymorphisms from the MMP genes and the survival outcomes in colorectal cancer. Hettiaratchi *et al.* (239) reported that one *MMP1* polymorphism (-1607 indelG in the promoter; NM_001145938.1:c.-1719delG) was associated with better OS, but this was not replicated in other studies (240-242). Langers *et al.* (243) reported the -1306C/T *MMP2* polymorphism; NG_008989.1:g.3726C>T) to be associated with better OS in colorectal cancer patients, which was not detected in a number of other studies (244, 245).

Based on both the small number of studies and polymorphisms investigated, as well as the conflicting results reported in literature, it can be concluded that the potential associations of VEGF, VEGFR and MMP polymorphisms with colorectal cancer patient prognosis is neither well-established nor well-studied.

1.10. SNP-based and haplotype-based genetic association studies

The human genome contains many sequence variations. These genetic variations include single nucleotide polymorphisms (SNPs), insertion/deletion of one or more nucleotides (indels), and microsatellite repeats (246). Of these, SNPs are the most frequent, with an estimated number of more than eleven million (246). SNPs occur within both coding and non-coding regions of genes and within intergenic regions. The SNPs in or close to genes can have functional consequences, such as changing amino acid sequences, affecting mRNA stability or altering gene expression levels (247).

Some of the variants in human DNA are the causes of the differences in phenotype and disease risks (246). There has been a major interest in identifying the genetic variations that can affect susceptibility to common diseases, and response to medical treatment (248-250). Thousands of GWAS have been published, some of which have identified common genetic variants conferring risk to specific diseases. For example, almost 4,000 SNP associations have been identified in ~200 diseases and traits (251). In these studies, usually the association of individual SNPs with the disease risk have been tested (SNP-based association studies).

Many researchers have suggested that haplotype analysis may provide additional information (252). Haplotypes are the combinations of alleles at different genomic loci. In some cases, haplotype analysis maybe more powerful than a SNP analysis, because the combination of several genetic variations may be associated with the phenotype (253-257). For a given genomic region on autosomal chromosomes, each individual inherits two sets of haplotypes, one from each parent (258). The commonly used haplotype phasing software include Arlequin (259, 260), PHASE II (261, 262), and Haplotyper (263). These applications can be used to predict the phased haplotypes of an individual by assigning the best possible combination of paired haplotypes based on the genotype data (261-263). The disadvantage of these statistical packages is that their results are not always accurate because a proportion of the inferred haplotypes may be incorrect (261-263). This is because it is often impossible to be certain about the haplotypes carried by one individual unless a family analysis is done (261-263).

While the genetic prognostic studies that test the association of genetic variations with survival outcomes of cancer patients is a relatively new field, both SNP-based and

haplotype-based association studies have been performed in colorectal cancer (14, 15). SNP-based and haplotype-based analyses can be complementary approaches in identifying the prognostic associations of genetic variations and genes in cancer.

1.11. Rationale, hypothesis and specific objectives of the research project

Rationale and hypothesis

Extensive biological and clinical findings suggest that abnormalities in angiogenesis, lymph-angiogenesis and metastasis may affect tumor progression and patient survival. Despite this strong evidence, the genetic basis of this relationship remains poorly characterized. In this study, I hypothesize that genetic alleles and their combinations as haplotypes from select genes acting in the angiogenesis, lymph-angiogenesis, and metastasis pathways are associated with clinical outcome in colorectal cancer patients.

Specific objectives

The overall aim of this research study is to identify new candidate markers that, once validated, may be used to improve prognostic accuracy in colorectal cancer patients. The specific objectives of this study are:

1. To investigate the associations between 381 individual genetic variants within 30 angiogenesis, lymph-angiogenesis and metastasis genes and outcome in a cohort of 505 colorectal cancer patients from NL.
2. To investigate the associations of haplotypes for these genes with outcome in the same patient cohort.

Chapter 2: Patient Cohort and Methodology

2.1. Ethics approval

This study was approved by the Health Research Ethics Board of Newfoundland (HREB Reference # 12.206).

2.2. Credits and collaborations

Lydia A. Dan: prepared the bfile to be used by the PLINK software to extract the genotypes and other related information for polymorphisms investigated in the study cohort; performed statistical analysis on the clinicopathological and treatment-related features and the 381 polymorphisms described in this thesis document; ran the PHASE II program together with Salem Werdyani using the input files prepared by Salem Werdyani; organized and interpreted the results with the help of the thesis supervisor; prepared the linkage disequilibrium map of the *MMP8-MMP27* genomic region; performed literature searches in order to interpret and discuss the results as described in this thesis document.

Salem Werdyani: prepared the input files for the PHASE II program and ran PHASE II to predict the haplotypes; involved in the preparation of the bfile to be used by the PLINK software to extract SNP genotypes and other relevant information.

Jingxiong Xu: from Princess Margaret Hospital, Toronto, Ontario; helped to perform quality control and population structure analyses based on the genotype data of the patient cohort.

Konstantin Shestopaloff: from University of Toronto, Ontario; contributed to the quality control and population structure analyses based on the genotype data of the patient cohort.

Dr. Patrick Parfrey: provided the genetic, clinicopathological and prognostic data used in this analysis.

Dr. Roger Green: provided the genetic, clinicopathological and prognostic data used in this analysis.

Dr. Wei Xu: from Princess Margaret Hospital, Toronto, Ontario; led the quality control and population structure analyses based on the genotype data of the patient cohort; helped with the study design, haplotype and statistical analyses and interpretation of the results.

Dr. Sevtap Savas: processed and coded the prognostic data for the patient cohort used in this study; combined the coded prognostic data with the coded genotype data for statistical analyses; provided the baseline characteristics tables as well as the statistical results on comparison of the NFCCR cohort (n=736) and the patient cohort investigated in this study (n=505); designed and led the project and supervised the thesis author throughout her program.

NFCCR Investigators: many investigators and personnel including Dr. Jane Green and Dr. Betty Dicks have contributed to the data collected and patients recruited to NFCCR. I gratefully acknowledge their contributions to this project.

Funding agencies

This study was supported by funds by the Canadian Institutes of Health Research (CIHR), Research and Development Corporation of Newfoundland (RDC), and the Medical Research Foundation (MRF) of Faculty of Medicine, Memorial University.

2.3. Patient cohort

A sub-cohort of patients recruited to the Newfoundland Colorectal Cancer Registry (NFCCR) was investigated in this study. The NFCCR was established in 1999 (264). Patients were eligible to join the NFCCR if they were diagnosed with colorectal carcinomas between 1 January 1999 and 31 December 2003 and were under 75 years of age. Informed consent was obtained from either the patients or their family proxies. In this cohort, there are 736 stage I-IV patients. These patients were followed up by the NFCCR until 2010. Collection of the prognostic data was described previously (265). Of these 736 patients, clinicopathological and prognostic data and DNA samples extracted from blood were available for 539 patients. These 539 patients were genotyped as described in **Section 2.4**. Out of 539, 505 patients were selected to be included in this study as described in **Section 2.5**.

The NFCCR also provided patient and disease related variables including age at diagnosis, sex, disease stage, tumor grade, vascular and lymphatic invasion status, tumor histology, MSI status, tumor location, familial risk status, *BRAF*-Val600Glu mutation status, adjuvant chemotherapy, adjuvant 5-Fluorouracil (FU)-based chemotherapy, and adjuvant radiotherapy status. Familial risk status was determined by NFCCR investigators using the Amsterdam and Bethesda criteria based on the patient family history as described in Green *et al.* (264). The MSI and *BRAF*-Val600Glu mutation status of the tumors were determined as described in Woods *et al*. (266).

2.4. Genotyping

Genomic DNA from 539 colorectal cancer patients (for whom prognostic data were available) was genotyped using the Illumina® human Omni1-Quad genome-wide SNP genotyping platform in an outsourced genomic facility (Centrillion Bioscience, USA). The chip used is designed based on tagSNP (i.e. tagging SNP) data and contains 1,134,514 SNP probes. The genomic coverage rate is about 93% and the median distance between the SNPs is 2.6 kb (267). Approximately 123,000 SNPs failed to be genotyped in this genotyping experiment. The genotypes of the remaining SNPs were recorded in a bfile (binary data file) by the outsourced genomics facility (Centrillion Bioscience, USA).

2.5. Quality control measures and inclusion-exclusion criteria for patients and genotype data

Quality control measures and inclusion-exclusion criteria were implemented on the data of 539 patients in order to have an ethnically homogenous population that consists of patients with high-quality genotype data. The following analyses were performed by Jingxiong Xu, Konstantin Shestopaloff and Dr. Wei Xu at the University of Toronto and the Princess Margaret Hospital, Toronto, Ontario. 1) Using the X-chromosome heterozygosity rate analysis, one sample was excluded from further analysis because the gender information indicated by the genetic data did not match the recorded gender of the patient. 2) The data was checked for individuals with a high missing genotype rate (>5%), but none of the patients failed this condition (i.e. all patients had >95% genotype call rates). 3) The data was checked for duplicate DNA samples but no

accidentally duplicated sample in the patient cohort was identified. 4) Among the 539 patients, 1st, 2nd and 3rd degree relatives who share similar genetic profiles were checked using the Identity by Descent method (268). As a result, a total of 21 patients (based on PI-Hat score threshold of >0.13) were excluded from our analysis. 5) Individuals with the outlying heterozygosity rate were identified using the mean heterozygosity rate information for each patient. As a result, one patient was excluded. 6) The patients' ethnicities were estimated with two statistical methods; multidimensional scaling (MDS; 269), and Principal Component Analysis (PCA; 270, 271). The public HapMap III Caucasian population data was used as a reference for the MDS analysis. As a result of these analyses, 11 samples were identified as population outliers (i.e. non-Caucasians). After this filtering, 505 patients met the quality control and inclusion-exclusion criteria and were included in the analysis. **Table 2.1** shows the baseline characteristics of the 505 colorectal cancer patients that constituted the study cohort.

2.6. Genes selected for this project

By literature search, 31 genes were identified that play biological roles in angiogenesis, lymph-angiogenesis or metastasis (**Table 2.2**) and were selected for this project.

Table 2.1: Baseline characteristics for the 505 patients included in this study

Variables	N	%
Sex		
Female	198	39.2
Male	307	60.8
Age at diagnosis	median: 61.43 years (range: 20.7-75)	
Histology		
non-mucinous	448	88.7
Mucinous	57	11.3
Location		
Colon	334	66.1
Rectum	171	33.9
Stage		
I	93	18.4
II	196	38.8
III	166	32.9
IV	50	9.9
Grade		
well/moderately differentiated	464	91.9
poorly differentiated	37	7.3
Unknown	4	0.8
Vascular invasion		
Absent	308	61
Present	159	31.5
Unknown	38	7.5
Lymphatic invasion		
Absent	298	59
Present	167	33.1
Unknown	40	7.9
OS status		
Alive	334	66.1
Dead	170	33.7
Unknown	1	0.2
OS follow up time	median: 6.36 years (range: 0.38-10.88)	
DFS status		
recurrence, metastasis or death (-)	304	60.2
recurrence, metastasis or death (+)	200	39.6
Unknown	1	0.2
DFS follow up time	median: 6 years (range: 0.22-10.88)	

Familial risk		
low risk	250	49.5
moderate/high risk	255	50.5
MSI status		
MSI-L/MSS	431	85.3
MSI-H	53	10.5
Unknown	21	4.2
Tumour *BRAF* Val600Glu mutation		
Absent	411	81.4
Present	47	9.3
Unknown	47	9.3
adjuvant chemotherapy status		
not given	224	44.4
Given	277	54.9
Unknown	4	0.79
adjuvant 5-FU based chemotherapy status		
not given	230	45.5
Given	261	51.7
Unknown	14	2.8
adjuvant radiotherapy status		
not given	364	72.1
Given	124	24.6
Unknown	17	3.4

OS: Overall Survival, DFS: Disease Free Survival, 5-FU: 5-Fluorouracil, MSI-H: microsatellite instability-high; MSI-L: microsatellite instability-low, and MSS: microsatellite stable

Table 2.2: Angiogenesis, lymph-angiogenesis and matrix metalloproteinase genes selected for this project and their genome coordinates

Gene symbol	Chromosome	Start (bp)	End (bp)	
VEGFA	6	43737946	43754223	VEGF Ligands
VEGFB	11	64002056	64006736	
VEGFC	4	177604691	177713895	
VEGFD	X	15363713	15402535	
PGF	14	75408533	75422467	
VEGFR1	13	28874483	29069265	VEGF Receptors
VEGFR2	4	55944426	55991762	
VEGFR3	5	180028506	180076624	
MMP1	11	102660641	102668966	MMPs
MMP2	16	55513081	55540586	
MMP3	11	102706528	102714342	
MMP7	11	102391239	102401478	
MMP8	11	102582526	102595685	
MMP9	20	44637547	44645200	
MMP10	11	102641233	102651359	
MMP11	22	24115036	24126503	
MMP12	11	102733464	102745764	
MMP13	11	102813721	102826463	
MMP14	14	23305793	23316803	
MMP15	16	58059282	58080804	
MMP16	8	89049460	89339717	
MMP17	12	132312941	132336316	
MMP19	12	56229214	56236767	
MMP20	11	102447566	102496063	
MMP21	10	127455027	127464390	
MMP23B	1	1567560	1570030	
MMP24	20	33814539	33864804	
MMP25	16	3096682	3110724	
MMP26	11	5009424	5013659	
MMP27	11	102562415	102576468	
MMP28	17	34092876	34122640	

The 23 MMPs listed above are the only MMPs in the human genome based on the information in the HUGO database (205).

2.7. Patient SNP genotype data

After the patients and genes to be included in this study were determined, the next step was to identify the SNPs to be investigated. We investigated SNPs irrespective of exonic or intronic locations along the genes. The genotyping platform annotates the genomic positions based on the human genome assembly hg19 (GRCh37). Hence, genome coordinates for each gene selected were retrieved using the UCSC genome browser (hg.19) (**Table 2.2**; 272, 273). If a gene had multiple transcripts, the genome coordinates of the longest isoform were retrieved so that all SNPs located in the gene region could be investigated. Using the genome coordinate information is a practical solution as by a single PLINK application (274, 275), patient genotype information for each SNP located within the genome coordinates (thus within the genes) could be retrieved from the bfiles.

Next, a new bfile was created using PLINK (274, 275). This bfile contained the genotype data of the 505 patients included in the analysis. In addition, as a quality control measure, SNPs whose genotype frequencies deviated from the Hardy-Weinberg Equilibrium (HWE; $p \leq 0.0001$) were excluded from this bfile. Also, this bfile only contained the SNPs with ≤5% missing genotype data as well as the SNPs with minor allele frequencies (MAFs) ≥5%. Once this new bfile was created, using the genome coordinates of genes as an input file, the genotype and other information related to SNPs were retrieved using PLINK (274, 275).

The number of SNPs for the 31 genes is shown in **Table 2.3**. For one of the MMP genes (*MMP23B*) there was no SNP genotype data in the patient cohort. Thus our final analysis included 381 SNPs in 30 genes (**Appendix A**).

Table 2.3: Number of SNPs in genes studied in this project

Gene	Number of SNPs
VEGFA	11
VEGFB	2
VEGFC	19
VEGFD	7
PGF	2
VEGFR1	49
VEGFR2	19
VEGFR3	20
MMP1	10
MMP2	22
MMP3	4
MMP7	5
MMP8	9
MMP9	6
MMP10	11
MMP11	3
MMP12	3
MMP13	4
MMP14	9
MMP15	4
MMP16	70
MMP17	13
MMP19	3
MMP20	21
MMP21	3
MMP23B	-
MMP24	25
MMP25	7
MMP26	1
MMP27	17
MMP28	2
Total = 31	**Total = 381**

Almost all of the variants were SNPs (n=380) while one was an insertion/deletion (indel). For simplicity, I refer to all of these variants as SNPs in this thesis. Each of the 381 SNPs was manually confirmed to be located in these genes using the dbSNP (276) and UCSC databases (272, 273). The PLINK extracted data were then processed in Microsoft® Excel for further analysis.

2.8. Variable coding and estimation of the best genetic model for each SNP

The variables for the clinicopathological, molecular, and treatment-related features (**Table 2.1**) were categorized as follows: sex (females=0, males=1), tumor histology (non-mucinous=0, mucinous=1), tumor location (colon=0, rectum=1), tumor stage (stage I=1, II=2, III=3, and IV=4), tumor grade (well or moderately differentiated=0, poorly differentiated=1), vascular invasion (absent=0, present=1), lymphatic invasion (absent=0, present=1), familial risk (low=0, high/intermediate=1), MSI status (MSS/MSI-L=0, MSI-H=1), *BRAF*-Val600Glu mutation status (wild-type=0, mutated=1), adjuvant chemotherapy given (no=0, yes=1), adjuvant 5-FU-based chemotherapy given (no=0, yes=1), and adjuvant radiotherapy (no=0, yes=1). Age was analyzed as a continuous variable.

The major allele (the more frequent allele) and the minor allele (the less frequent allele) for each SNP were determined based on the patient cohort genotype data. The genotype data obtained was coded in several different ways depending on the purpose. Traditionally, in the absence of information on the true underlying genetic model, the effects of polymorphisms on outcome is investigated by using one or more of the four

genetic models: additive, co-dominant, dominant and recessive. These genetic models are described elsewhere (277). Briefly, assuming the models are modeled based on the minor alleles (the least frequent allele in the patient cohort); in the dominant genetic model the survival times of patients with the homozygous minor allele (aa) and heterozygous genotypes (Aa) are compared with the survival times of patients with the homozygous major allele genotype (AA). In the recessive genetic model, the survival times of patients with the homozygous minor allele genotype (aa) are compared with the survival times of patients with the homozygous major allele (AA) or heterozygous (Aa) genotypes. In the co-dominant model, the survival times of patients with the heterozygous (Aa) and homozygous minor allele genotypes (aa) are compared separately to survival times of patients with the homozygous major allele genotype (AA). In the additive genetic model, the survival times of patients with the homozygous minor allele (aa), heterozygous (Aa) and homozygous major allele (AA) genotypes are analyzed simultaneously as a continuous variable.

For this project, I applied a previously published strategy to estimate the best genetic model for each SNP using the Kaplan Meier survival curves constructed assuming the co-dominant genetic model (277). The main advantage of this strategy is that it helps estimates the best genetic model for each SNP based on their characteristics, rather than applying one or more genetic models randomly to the whole set of polymorphisms (277). There are other ways to determine the best genetic models for SNPs. For example, the SNP data can be investigated for each of the genetic models by separate univariable Cox regression analysis and the genetic model with the lowest p-value can be deemed to be

the best fitting genetic model (278). However, this approach creates a multiple testing issue because of large number of tests performed (277).

In this study, first, the patient genotypes were coded assuming the co-dominant genetic model (or additive; both coding are identical) by using a PLINK command. Kaplan Meier survival analysis (279) was performed for each of the 381 SNPs to choose the genetic model that best fit each SNP. This analysis was done separately for OS and DFS. The Kaplan Meier survival curves were then inspected by two individuals (the author and supervisor). By looking at the pattern of the curves, one can estimate which genetic model or models (dominant, recessive, co-dominant or additive) may best fit the genotypes of a polymorphism. When in doubt, multiple genetic models were chosen. In cases where Kaplan Meier curves did not separate well or clear enough for us to estimate a genetic model, polymorphism were excluded from further statistical analysis. I examined the SNPs with the number of aa genotype <10 using the dominant genetic model. Results of this analysis are summarized in **Table 2.4**.

After this step, genotypes were re-coded using a Microsoft® Excel function for the genetic model assigned to each SNP. The genotype data were then combined with the clinicopathological, demographic, molecular and prognostic data of the patients in Microsoft Excel® sheets. The files were then imported into IBM SPSS software (v.19 and v.20) for statistical analysis.

Table 2.4: Summary of the best genetic models predicted for the 381 SNPs

Estimated genetic model	Number of SNPs	
	OS	DFS
Recessive only	137	136
Dominant only	104	103
Co-dominant only	29	41
Additive only	0	0
Multiple genetic models	20	29
*Excluded	91	72
Total	**381**	**381**

*SNPs excluded from further analysis when their Kaplan Meier curves did not separate clear enough to estimate a genetic model.

2.9. Gene-based haplotype survival analysis

In order to perform the haplotype-survival association analysis, the phased haplotypes for each gene in each patient were estimated using PHASE II (v.2.1.1) software (261, 262). PHASE II also estimated the haplotype frequencies.

In brief, PHASE II software was downloaded from the site of University of Chicago (www. http://stephenslab.uchicago.edu/software.html#phase) (280). Input text files that contained the SNP genotype data of the patient cohort were created for each gene separately using Perl programs written by Salem Werdyani. Input files were created for 29 genes, as the *MMP26* gene had a single SNP genotyped and thus haplotype estimation was not relevant. Then using the PHASE II commands, the phased haplotypes

for each patient were estimated. To increase the accuracy of predictions, the estimations were performed for five rounds as recommended by the PHASE II developer.

For the X-chromosome-linked genes (e.g. *FIGF* in this project), the PHASE II input files were created differently as recommended by the software developer. For *FIGF*, the males were paired separately in a file and assigned as "known individuals" (as males have one X chromosome, their X-linked haplotypes are easily deducible). In contrast, the female individuals were paired and assigned as "unknown individuals". Preparation of input files and estimation of haplotypes for this gene were then preceded as explained above.

After the phased haplotypes were estimated for each gene, the haplotypes together with haplotype frequency information generated by PHASE II were combined in Microsoft® Excel files. In the survival analysis, survival of the patients with either one or two copies of the most frequent haplotype for each gene was compared with the survival of patients with the remaining haplotypes. I limited this study to the genes that had at least one haplotype in ≥ 5% of the patients. As a result, two genes, *VEGFR1* and *MMP16*, which did not have a frequent haplotype, were excluded. For *FIGF*, which is an X-linked gene, patients with either one copy (all males and females with one copy of the most frequent haplotype) or two copies (females only) were categorized together and compared with the patients with other haplotypes. For the haplotype-based analysis of the remaining genes, since the effect of the most common haplotype in homozygous and heterozygous state when compared to the remaining haplotypes (named as "other haplotypes throughout the thesis document) can be different from each other, similar to the SNP

analysis, I first estimated the best genetic model describing the effect of haplotype variables using the Kaplan Meier curves (**Section 2.8**).

Table 2.5 shows the number of different haplotypes estimated for each gene in this analysis. **Table 2.6** and **Table 2.7** summarize the best genetic model estimated for the haplotype variables in each gene for OS and DFS, respectively.

2.10. Measures of outcome

Overall survival (OS) was analyzed using the OS status and OS time (the time from diagnosis until the time of death from any cause). Disease-free survival (DFS) was analyzed using the DFS status and DFS time (time from diagnosis to the time of recurrence, metastasis or death from any cause). When a patient did not experience these events, they were censored at the date of the last follow-up.

2.11. Univariable survival analyses

The purpose of univariable analyses is to test for association between a variable (such as a genotype or a baseline variable) and the outcome of interest (in this case, overall or disease-free survivals). In this study, Kaplan-Meier survival and Cox regression methods were used for univariable survival analyses. The Kaplan-Meier curves show patients' survival characteristics and were used to select the best genetic model for each SNP and haplotype variables. On the other hand, the univariable Cox-regression analysis

Table 2.5: Number of phased haplotypes predicted for each gene

Gene	Number of common haplotypes (frequencies ≥ 5%)
VEGFA	7
VEGFB	3
VEGFC	5
VEGFD	4
PGF	2
**VEGFR1*	none
VEGFR2	2
VEGFR3	5
MMP1	7
MMP2	6
MMP3	5
MMP7	4
MMP8	5
MMP9	4
MMP10	7
MMP11	3
MMP12	4
MMP13	3
MMP14	8
MMP15	3
**MMP16*	none
MMP17	3
MMP19	3
MMP20	8
MMP21	3
MMP24	5
MMP25	5
MMP27	6
MMP28	3

*****There were no haplotypes estimated with frequencies ≥ 5% in these two genes (*VEGFR1* and *MMP16*). These genes therefore were not investigated during the haplotype association analysis. *MMP26*, for which only one SNP was investigated, is not included in this table.

Table 2.6: The best genetic models predicted for haplotype-based variables (overall survival)

Genes	Best predicted model for each gene (haplotypes)
VEGFA	Recessive
VEGFB	*Excluded
VEGFC	Dominant
VEGFD	*Excluded
PGF	Co-dominant
FLT4	Dominant
KDR	Dominant
MMP1	Recessive
MMP2	Recessive
MMP3	Recessive
MMP7	Dominant
MMP8	Recessive
MMP9	*Excluded
MMP10	Recessive
MMP11	Dominant
MMP12	Recessive
MMP13	*Excluded
MMP14	Dominant
MMP15	Co-dominant, dominant
MMP17	Recessive, co-dominant
MMP19	Additive, recessive, co-dominant, dominant
MMP20	Dominant
MMP21	Recessive
MMP24	Recessive
MMP25	Co-dominant
MMP27	Recessive
MMP28	Dominant

*Excluded from statistical analyses as the Kaplan Meier curves of the haplotype variables did not separate clear enough to predict a genetic model.

Table 2.7: The best genetic models predicted for haplotype-based variables (disease-free survival)

Genes	Best predicted model for each gene (haplotypes)
VEGFA	Recessive
VEGFB	Dominant
VEGFC	Co-dominant, recessive
VEGFD	*Excluded
PGF	Recessive
FLT4	Dominant
KDR	Dominant
MMP1	Co-dominant
MMP2	Recessive
MMP3	Recessive
MMP7	Dominant
MMP8	Recessive
MMP9	*Excluded
MMP10	Recessive, dominant
MMP11	Dominant
MMP12	Recessive
MMP13	Co-dominant
MMP14	Dominant
MMP15	Recessive
MMP17	Recessive
MMP19	Additive, recessive, co-dominant, dominant
MMP20	Dominant
MMP21	Recessive
MMP24	Recessive
MMP25	Co-dominant
MMP27	Recessive
MMP28	Dominant

*Excluded from statistical analyses as the Kaplan Meier curves of the haplotype variables did not separate clear enough to predict a genetic model.

estimates a *p* value and the hazard ratio (HR) with 95% confidence intervals (CIs) (281). Univariable Cox regression analysis was performed for those SNPs and haplotypes for which a genetic model was chosen based on the Kaplan Meier curves.

Univariable analysis was also performed for the clinicopathological, molecular and treatment-related variables in order to identify the baseline variables that would be entered into the multivariable model, together with the SNP genotypes or haplotypes that met the significance threshold requirements. **Appendix B** and **Appendix C** show the univariable Cox regression analysis results for these variables for overall survival and disease free survival, respectively. The significance threshold set for the baseline variables as well as the haplotype-based analysis was $p<0.05$. Due to the large number of polymorphisms investigated, to account for multiple testing while also limiting the false-negative associations (i.e. when there is a real association, which is missed because of a conservative multiple testing correction), the significance threshold for association of the polymorphisms was set at $p<0.001$ prior to statistical analysis. Univariable analyses were conducted using IBM SPSS (version 19 and 20). The results were then exported from IBM SPSS into Microsoft® Excel sheets.

2.12. Identification of highly correlated variables

Spearman's correlation test was used to check whether the variables investigated were highly correlated. Variables with a correlation score (r_s) ≥ 0.8 were deemed to be highly correlated. This information is very important because multicollinearity in a model may inflate the standard errors, thus making some variables appear statistically

insignificant while they should be significant or vice versa (282). To avoid this situation, one of the correlated variables should be excluded from the final multivariable model.

This test was performed for the baseline, molecular and treatment-related characteristics as well as for the SNPs that were significantly associated with outcomes in the univariable analysis. Based on the results of this test, only lymphatic and vascular invasion ($r_s = 0.963$), and adjuvant chemotherapy and adjuvant 5-FU-chemotherapy status ($r_s = 0.992$) were highly correlated with each other. Among these variables, I reasoned that the one with the smallest p-value in the univariable analysis and with less missing data should be included in the baseline multivariable (MVA) model. On this basis, vascular invasion and adjuvant 5-FU-based chemotherapy were included in the baseline models. Of note, none of the SNPs that were significantly associated with outcome in the univariable analysis was associated with these baseline variables.

2.13. Multivariable Cox regression analysis

Multivariable Cox regression analysis assesses whether several covariates independently influence outcome, i.e. it shows the independent predictive potential of each variable in a model (282). I performed this analysis for three purposes: a) to identify the baseline variables that would be included in the final multivariable models, b) to construct the final multivariable models containing both the baseline variables and the SNPs, and c) to construct the multivariable models containing both the baseline variables and the haplotypes.

In order to get our baseline model, the clinicopathological, molecular, and treatment-related baseline variables with $p < 0.05$ in the univariable Cox regression

analysis were entered into multivariable models for OS and DFS separately. As explained in **Section 2.12**, in the case of variables that were highly correlated with each other, one of them was excluded from this analysis. The baseline variables that remained significant after this analysis were selected to enter the final multivariable model together with the polymorphisms or haplotype variables significantly associated with outcomes in the univariable analyses. As a result, stage and MSI status remained significant for both overall and disease-free survival in the multivariable models. Age was not significant in our analysis, but since it is a well-established prognostic marker, especially in overall survival, I opted to construct our final multivariable models both with and without age as a covariate. These analyses were conducted using the IBM SPSS (version 19 and 20). The results were then exported from IBM SPSS and organized in Microsoft Excel® spread sheets.

2.14. Construction of linkage disequilibrium map of the genomic region encompassing the *MMP8* and *MMP27* genes

A linkage disequilibrium (LD) map of the genomic region containing the *MMP8* and *MMP27* genes was constructed using Haploview 4.2 software (283). In order to construct the LD map, the genotypes for the polymorphisms located within the genomic region of these two genes were first extracted using PLINK. These data were then formatted and used in the Haploview to visualize the LD map of the region.

Chapter 3: Results

3.1. Univariable survival analyses

3.1.1 Single SNP survival association analysis

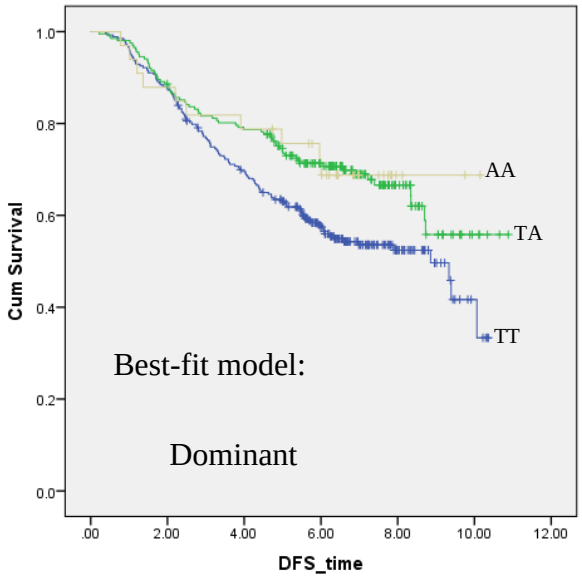
In this study, 381 polymorphisms genotyped in a cohort of 505 colorectal cancer patients were investigated for their associations with survival outcomes. Kaplan Meier curves were constructed in order to choose the genetic model that best fits each SNP (277). **Figure 3.1** shows examples of the Kaplan Meier curves constructed for this purpose. As a result of this analysis, I was able to choose the best genetic model(s) for 290 and 309 SNPs for overall survival and disease-free survival, respectively (**Table 2.4**). These SNPs were then investigated in a Cox univariable analysis.

Polymorphisms associated with overall survival

Of the polymorphisms investigated, three SNPs were found to be significantly associated with overall survival and all of these associations were observed under the dominant genetic model. The results of univariable Cox regression analysis for these SNPs and overall survival are summarized in **Table 3.1**. The minor allele was protective in all three SNPs (minor allele frequency ~ 27%). For the *MMP8*-rs12365082 (NM_002424.2:c.*1247A>T) polymorphism, I observed that patients with the TA or AA genotypes were at lower risk of death compared to patients with the TT genotype (**Table 3.1**; **Figure** 3.2.**a**).

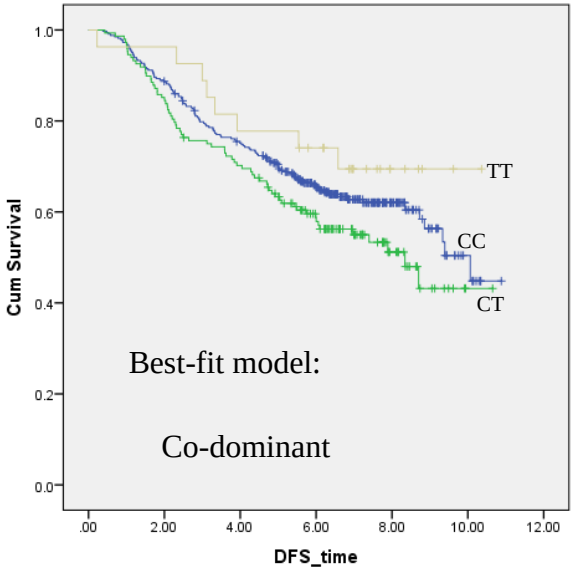
Figure 3.1: Examples of Kaplan-Meier survival plots

MMP8-rs12365082 T/A



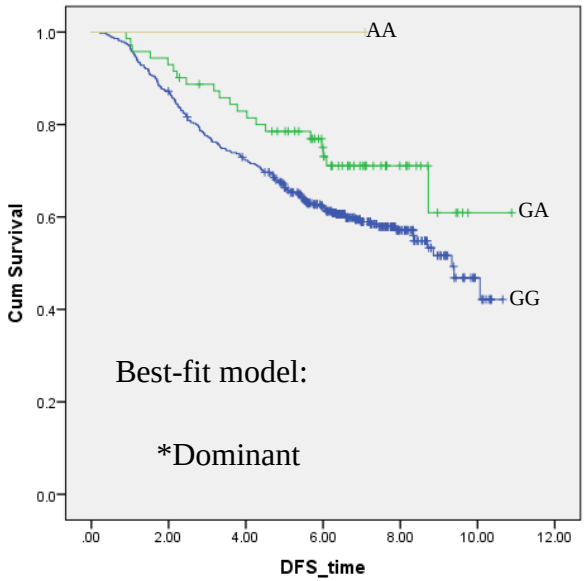
Minor allele: A

MMP3-rs3020919 C/T



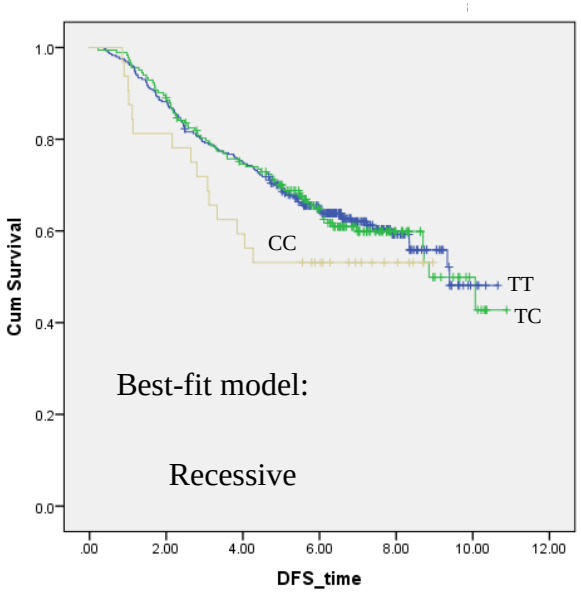
Minor allele: T

MMP1-rs10488 G/A



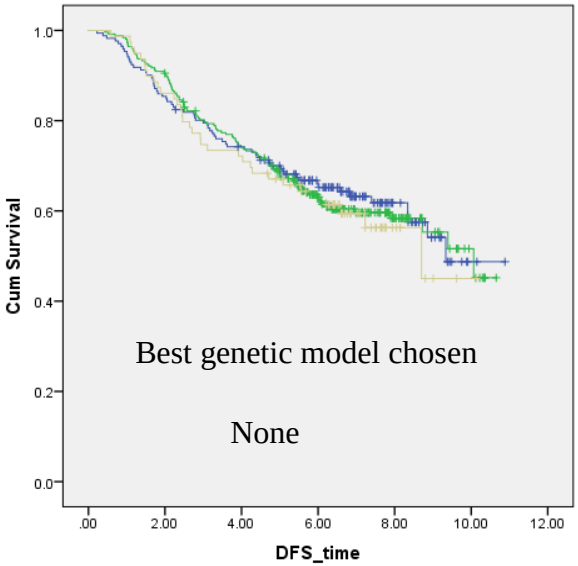
Minor allele: A

MMP8-rs2012390 T/C



Minor allele: C

MMP16-rs3851539 A/G



Minor allele: G

Blue = major allele homozygotes, green = heterozygotes, beige = minor allele homozygotes. *An example of a case where the dominant genetic model was chosen by default as the number of patients with the minor allele homozygote genotype (aa) was less than 10.

Table 3.1: Polymorphisms associated with overall survival in the univariable analysis (dominant genetic model) (n= 504)

SNP	Genotype categories	p-value	HR	95% CI for HR		Minor allele (MAF)
				Lower	Upper	
MMP8-rs12365082	TA +AA vs TT	**0.0006**	0.579	0.423	0.791	A (0.2683)
MMP27-rs11225388	AG + GG vs AA	**0.0005**	0.574	0.42	0.785	G (0.2693)
MMP27-rs11225389	CA + AA vs CC	**0.0005**	0.574	0.42	0.785	A (0.2693)

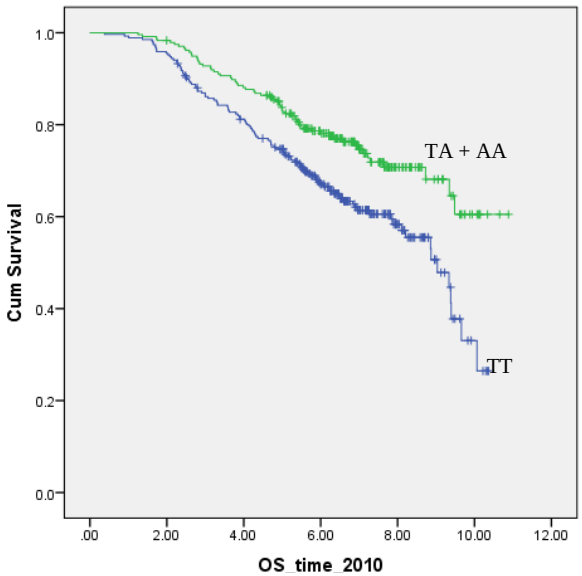
HR: hazard ratio, CI: confidence interval

For the *MMP27*-rs11225388 (NM_022122.2:c.103-233T>C) polymorphism, patients with AG or GG genotypes had a longer overall survival than those with the AA genotype (**Table 3.1**; **Figure** 3.2.**b**). Finally, patients carrying the CA or AA genotypes of the *MMP27*-rs1225389 (NC_000007.14:g.155851799T>A) polymorphism had longer overall survival than those homozygous for the major allele (CC genotype) (**Table 3.1**; **Figure** 3.2.**c**). None of the remaining SNPs tested were associated with overall survival at the significance threshold of $p = 0.001$.

Upon further investigation, I found that the genotypes of these three SNPs were highly correlated with each other (Spearman's correlation coefficient r_s values: between *MMP8*-rs12365082 and *MMP27*-rs11225388 = 0.997, *MMP8*-rs12365082 and *MMP27*-rs1225389 = 0.996, and between *MMP27*-rs11225388 and *MMP27*-rs1225389 = 0.999). The *MMP8* and *MMP27* genes are close to each other on chromosome 11 where there is a cluster of nine MMP genes (**Figure 3.3**).

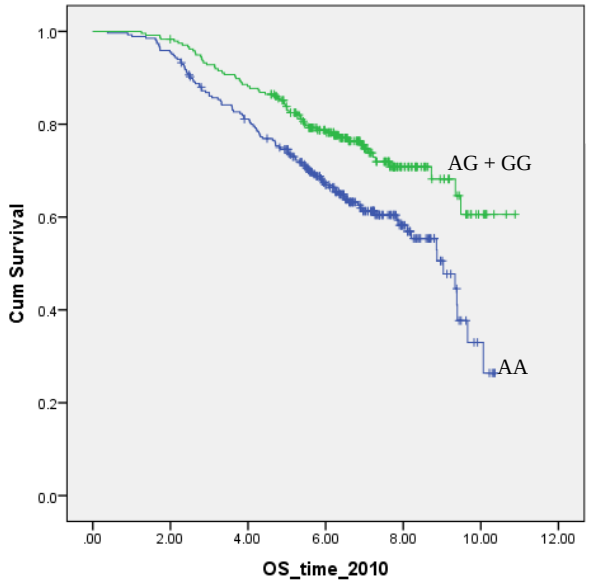
Figure 3.2: Kaplan-Meier survival plots for the three polymorphisms associated with overall survival

a. *MMP8*-rs12365082 T/A



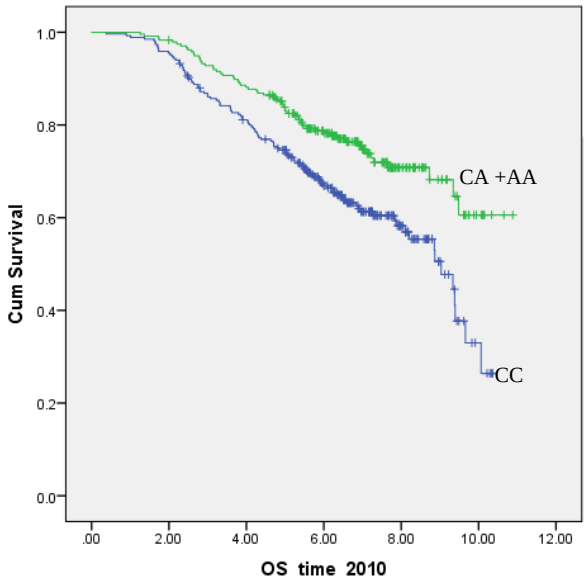
Minor allele: A

b. *MMP27*-rs11225388 A/G



Minor allele: G

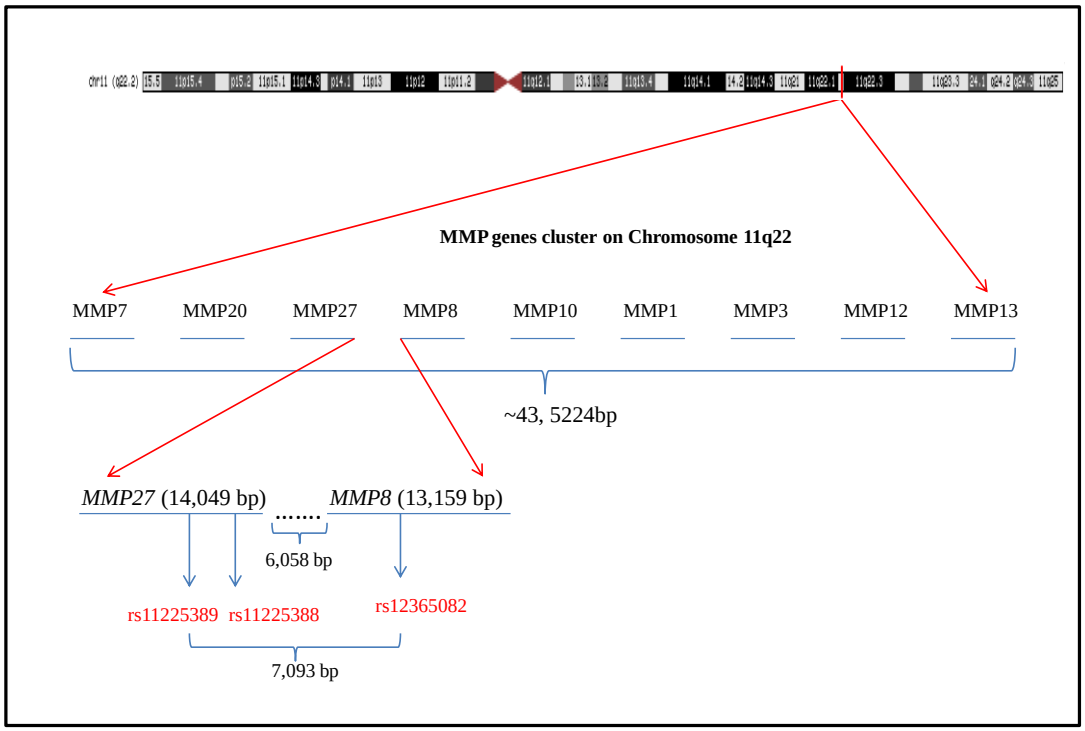
c. *MMP27*-rs11225389 C/A



Minor allele: A

Blue = major allele homozygous genotype, green = heterozygous and homozygous minor allele genotypes

Figure 3.3: The MMP gene cluster on chromosome 11q22.



The *MMP27* and *MMP8* genes are 14,049 and 13,159 base pairs long, respectively. The distance between the *MMP27*-rs11225388 and *MMP8*-rs12365082 polymorphisms is 7,093 base pairs. Figure not drawn to scale. Chromosomal bar is obtained from the UCSC genome browser website (272, 273).

Polymorphisms associated with disease-free survival

In the univariable analysis, none of the polymorphisms investigated in this study were found to be associated with disease-free survival in our patient cohort at the significance threshold of p= 0.001.

3.1.2 Haplotype-based survival analysis

We performed haplotype survival association analyses in relation to both overall and disease-free survivals. Univariable survival analysis was performed for the haplotypes estimated for 27 genes. One gene without multiple polymorphisms (*MMP26*) and two genes without common haplotypes with frequencies ≥5% (*VEGFR1* and *MMP16*) were excluded from the haplotype analysis. The purpose of the analysis was to compare survival times of patients with one or two copies of the most frequent haplotype with the survival times of patients with the remaining haplotypes (**Section 2.9**). Similar to the approach used for SNP associations, Kaplan Meier curves were constructed to select the genetic model that best fit each haplotype category (**Section 2.8**). Haplotypes not distinguished by the Kaplan Meier curves (n = 4 for overall survival and n = 2 for disease-free survival) were excluded from further analysis (**Table 2.6** and **Table 2.7**). The remaining haplotypes (n = 23 for overall survival and n = 25 for disease free survival) were further investigated by Cox univariable analysis.

Haplotypes associated with overall survival

Haplotypes of three genes were associated with overall survival under the recessive (**Table 3.2**; *MMP3, MMP27*) or co-dominant (**Table 3.2**; *MMP25*) genetic models. For *MMP3*, patients homozygous for the most common haplotype had longer survival than patients with one or no copies of the most common haplotype (**Table 3.2**; **Figure** 3.4.**a**). An increased hazard was observed in patients homozygous for the most

common *MMP27* haplotype compared to those with a single copy of the most common haplotype or patients with other haplotypes (**Table 3.2**; **Figure** 3.4.**b**).

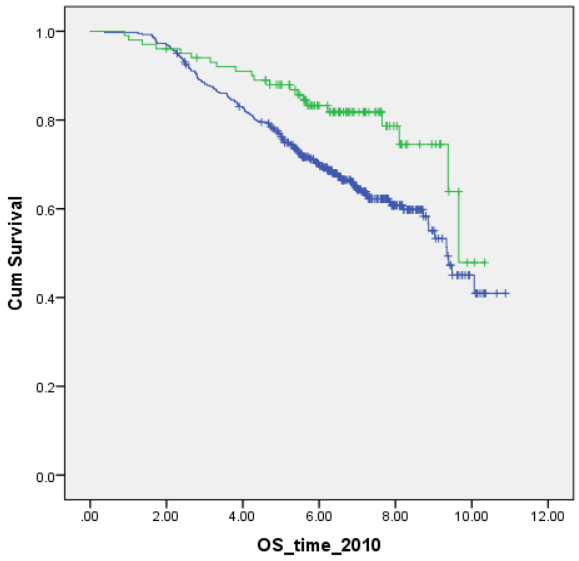
Table 3.2: Haplotypes associated with overall survival in a univariable survival analysis (n= 504)

			95% CI for HR		
Variable	**p-value**	**HR**	**Lower**	**Upper**	**Genetic model**
**MMP3* haplotype	**0.007**	0.533	0.337	0.842	Recessive
**MMP27* haplotype	**0.03**	1.523	1.041	2.228	Recessive
MMP25 haplotype	**0.032**				Co-dominant
***MMP25* haplotype	**0.009**	1.518	1.109	2.078	
****MMP25* haplotype	0.651	1.146	0.635	2.065	

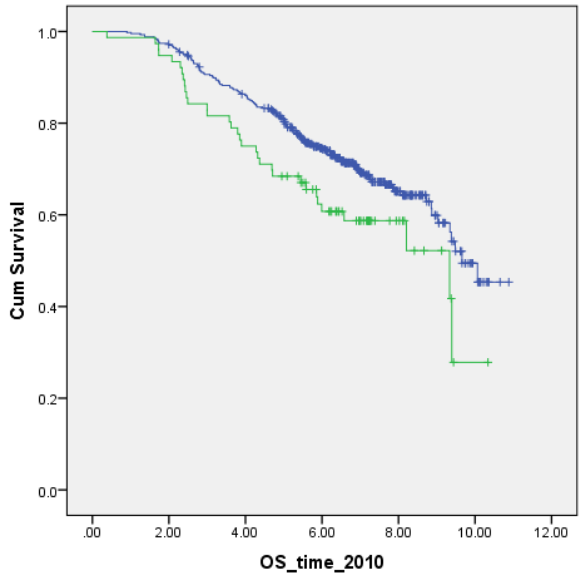
HR: hazard ratio, CI: confidence interval
*****patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes
******patients heterozygous for the most common haplotype vs patients with other haplotypes
*******patients homozygous for the most common haplotype vs patients with other haplotypes

Figure 3.4: Kaplan-Meier survival plots for the haplotypes associated with overall survival

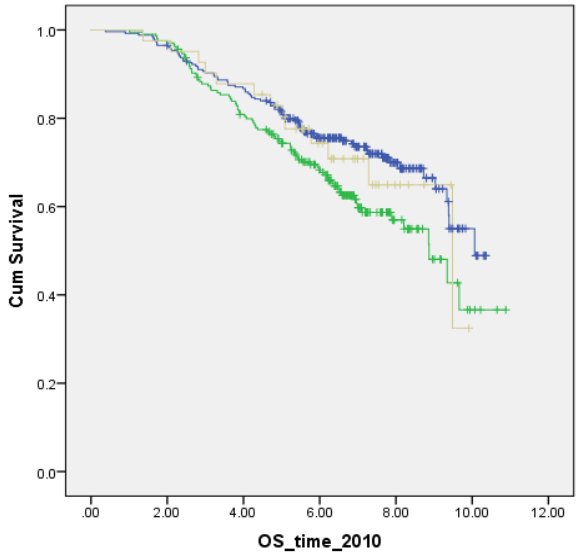
a. *MMP3*-haplotype



b. *MMP27*-haplotype



c. *MMP25*-haplotype



a) The best-fit model: recessive. Blue = other haplotypes or one copy of the most common haplotype. Green = two copies of the most common haplotype.
b) The best-fit model: recessive. Blue = other haplotypes or one copy of the most common haplotype. Green = two copies of the most common haplotype.
c) The best-fit model: co-dominant. Blue = other haplotypes. Green = one copy of the most common haplotype (heterozygotes). Beige = two copies of the most common haplotype (homozygotes).

Finally, I found that patients heterozygous for the most frequent *MMP25* haplotype had a higher risk of death compared to patients with other haplotypes (**Table 3.2**; **Figure** 3.4.**c**). Of note, there was no association between homozygosity for the most common *MMP25* haplotype and overall survival (**Table 3.2**).

The most common haplotypes of the *MMP3, MMP25,* and *MMP27* genes were quite frequent in the patient cohort (**Table 3.3**). The polymorphisms and alleles that constituted the most common haplotypes in these genes are shown in **Table 3.4**. The haplotypes consisted of four SNPs in *MMP3,* seven SNPs in *MMP25* and 17 SNPs in *MMP27*.

Table 3.3: Frequencies of the most common haplotypes for the five genes associated with survival in univariable analyses

Genes	Frequency	Survival times Associated
MMP3	44.16	OS, DFS
MMP25	28.71	OS, DFS
MMP27	39.96	OS, DFS
MMP8	44.88	DFS
MMP21	45.77	DFS

OS: overall survival, DFS: disease-free survival

Table 3.4: The most common haplotypes (frequency ≥ 5%) for the three genes associated with overall survival in univariable analyses

Gene	Haplotype	Frequency
MMP3	CACA	0.441604
MMP27	CCGTAAAACCAAAGAGC	0.399644
MMP25	TCGCTGC	0.287138

The rs numbers for the SNPs in each haplotype (starting with the SNP with the smallest genome coordinate along the chromosome where the gene is located to the SNP with the largest) is;

***MMP3*:** rs566125**,** rs3025066, rs3020919 and rs679620

***MMP27*:** rs2509010, rs11607205, rs1276289, rs11821641, rs1276286, rs2846723, rs2846701, rs2846703, rs3809018, rs4754870, rs17099425, rs11225386, rs11225388, rs2846707, rs1939015, rs12099177 and rs11225389

MMP25: rs2247226, rs10431961, rs7199221, rs1064875, rs1064948, rs11864930 and rs10438593

Haplotypes associated with disease-free survival

The results obtained in the univariable analysis are summarized in **Table 3.5**. Five genes were associated with disease-free survival: four under a recessive model (*MMP3*, *MMP8*, *MMP21*, and *MMP27*) and one under a co-dominant model (*MMP25*) (**Figure 3.5**).

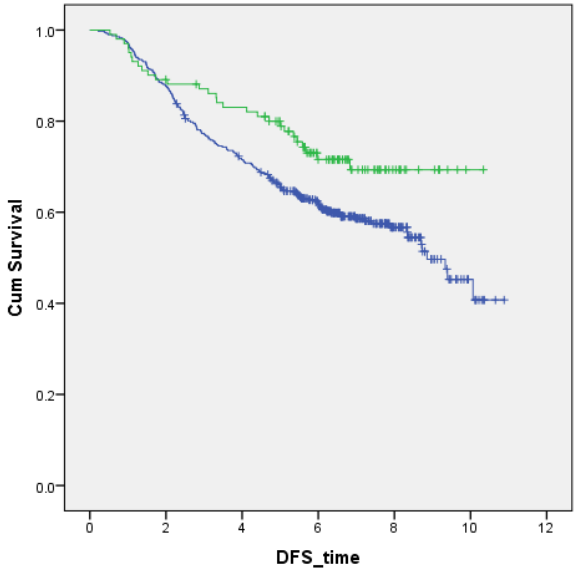
Table 3.5: Haplotypes associated with disease-free survival in univariable survival analyses (n= 503)

Variable	p-value	HR	95% CI for HR		Genetic model
			Lower	Upper	
**MMP3* haplotype	**0.021**	0.625	0.419	0.932	Recessive
**MMP8* haplotype	**0.01**	1.521	1.103	2.095	Recessive
**MMP21* haplotype	**0.032**	0.657	0.448	0.964	Recessive
**MMP27* haplotype	**0.027**	1.484	1.046	2.107	Recessive
MMP25 haplotype	0.121				Co-dominant
***MMP25* haplotype	**0.048**	1.338	1.002	1.786	
****MMP25* haplotype	0.964	0.987	0.563	1.732	

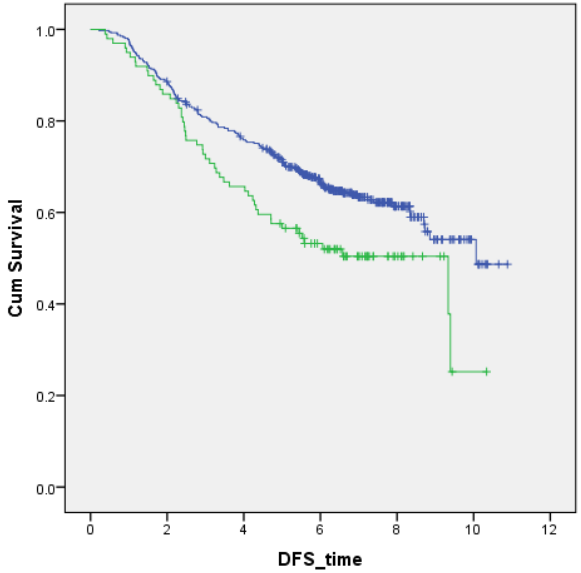
HR: hazard ratio, CI: confidence interval
*****patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes
****** patients heterozygous for the most common haplotype vs patients with other haplotypes;
*** patients homozygous for the most common haplotype vs patients with other haplotypes

Figure 3.5: Kaplan-Meier survival plots for the haplotypes associated with disease-free survival

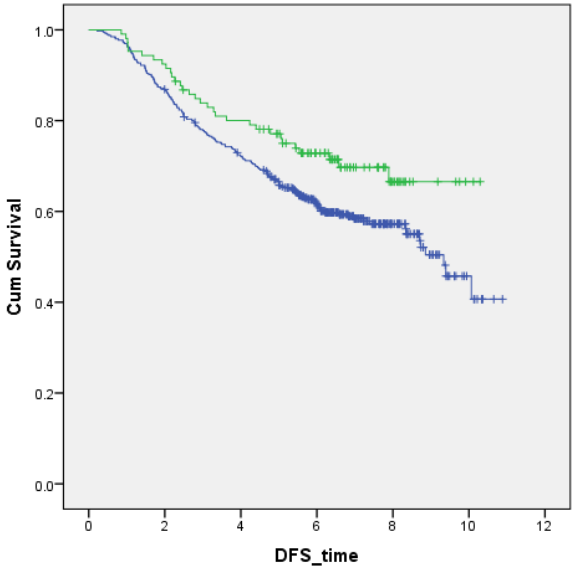
a. *MMP3*-haplotype



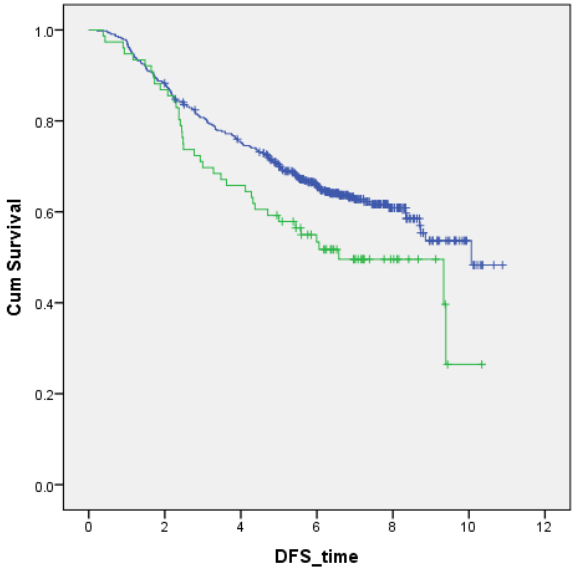
b. *MMP8*-haplotype



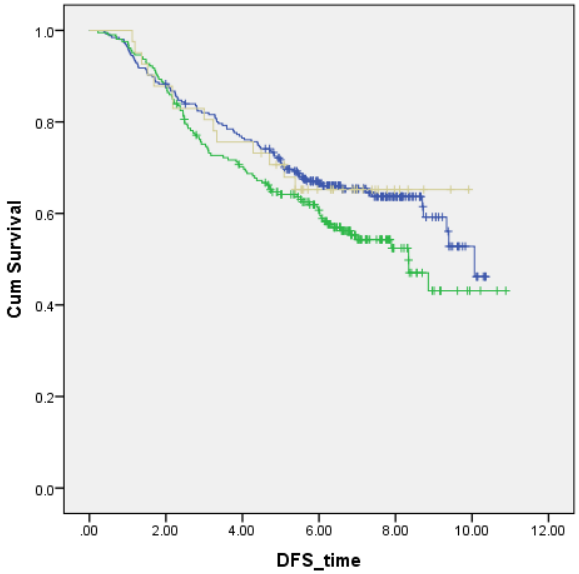
c. *MMP21*-haplotype



d. *MMP27*-haplotype



e. *MMP25*-haplotype



a) The best-fit model: recessive. Blue = patients with other haplotypes or one copy of the most common haplotype. Green = patients with two copies of the most common haplotype.
b) The best-fit model: recessive. Blue = patients with other haplotypes or one copy of the most common haplotype. Green = patients with two copies of the most common haplotype.
c) The best-fit model: recessive. Blue = patients with other haplotypes or one copy of the most common haplotype. Green = patients with two copies of the most common haplotype.
d) The best-fit model: recessive. Blue = patients with other haplotypes or one copy of the most common haplotype. Green = patients with two copies of the most common haplotype.
e) The best-fit model: co-dominant. Blue = patients with other haplotypes. Green = patients with one copy of the most common haplotype (heterozygotes). Beige = patients with two copies of the most common haplotype (homozygotes).

Patients homozygous for the most common *MMP3* haplotype had a 37% reduced risk of recurrence, metastasis or death when compared to other patients (**Figure** 3.5.**a**). Patients homozygous for the most common *MMP8* haplotype had a greater risk of disease recurrence, metastasis or death when compared to other patients (**Table 3.5**; **Figure** 3.5**b**). Patients homozygous for the most common *MMP21* haplotype had a 34% reduced risk of event when compared to patients with a single copy of the most common haplotype or patients with other haplotypes (**Table 3:5**; **Figure** 3.5**c**). Patients homozygous for the most common *MMP27* haplotype had a higher risk of recurrence, metastasis or death compared to other patients (**Table 3.5**; **Figure** 3.5**d**). Finally, in the case of the *MMP25* gene, patients who were heterozygous for the most common haplotype had decreased disease-free survival times (**Table 3.5**; **Figure** 3.5**e**) when compared to patients with other haplotypes. Of note, the associations of *MMP3*, *MMP27*, and *MMP25* haplotypes with disease-free survival were also observed in the overall survival analysis as described previously. The most common haplotype for each of these three genes is shown in **Table 3.4**. **Table 3.6** shows the most common haplotypes for *MMP8* and *MMP21* (frequencies in the study cohort are shown in **Table 3.3**).

Table 3.6: The most common haplotypes (frequency ≥ 5%) for the genes associated with disease-free survival in univariable analyses

Gene	Haplotype	Frequency
MMP8	TGCGTCCAG	0.45
MMP21	GTG	0.46

The rs numbers for the SNPs in each haplotype (starting with the SNP with the smallest genome coordinate along the chromosome where the gene is located to the SNP with the largest) is;

***MMP8*:** rs12365082, rs7934972, rs12284255, rs3740938, rs2012390, rs1940475, rs6590984, rs3765620 and rs2155052

***MMP21*:** rs7922546, rs10901424 and rs12775804

3.1.3 Survival analyses for baseline variables

Univariable analyses were performed to determine associations between baseline clinicopathological, molecular and treatment-related characteristics and survival times. The results were used to identify the variables to be included in the final multivariable models, together with the SNPs and haplotypes that were associated with overall or disease-free survivals in the univariable analyses.

Baseline variables associated with overall survival

Appendix B shows the results of a univariable Cox regression analysis for the baseline variables and overall survival. Of 14 baseline variables tested in the univariable analysis, five were associated with overall survival (sex, stage, vascular invasion,

lymphatic invasion and MSI). **Appendix D** shows the Kaplan Meier curves for these variables. As expected, male patients had a significantly higher risk of death than did females. Patients with stage III and stage IV disease had increased risks of death compared to those with stage I disease. Patients with vascular or lymphatic invasions of the tumour showed significantly greater hazard of death than did patients with no vascular or lymphatic invasions. Finally, patients with MSI-H tumor status had a lower risk of death than did patients with MSS or MSI-low tumors.

Baseline variables associated with disease-free survival

The results of a univariable Cox regression analysis for the baseline clinicopathological, molecular, and treatment-related variables and disease-free survival are summarized in **Appendix C**. Six variables were associated with disease-free survival as expected (sex, stage, location, vascular invasion, lymphatic invasion, and MSI status). Male patients had greater risks of disease recurrence, metastasis or death compared to the female patients. Patients with rectal cancer had shorter disease-free survival times compared to those with colon cancer. Shorter disease-free survival times were also observed in stage III and stage IV patients compared to stage I patients. Patients with vascular or lymphatic tumor invasion showed higher risk of disease recurrence, metastasis or death than patients with tumors lacking vascular or lymphatic invasion. Finally, patients having MSI-H tumors had reduced risk of events (disease recurrence, metastasis or death) when compared to patients with MSS or MSI-low tumors. Kaplan Meier curves for the variables significantly associated with disease free survival are shown in **Appendix E**.

3.2. Multivariable survival analysis

Selection of baseline covariates for the final multivariable models is described in **Section 2.13**. **Appendix F** and **Appendix G** show the results of a baseline multivariable Cox regression analysis results for overall survival and disease-free survival, respectively. The SNPs and haplotypes that met the significance threshold in the univariable analysis were entered into separate multivariable models together with the selected baseline variables, namely stage and MSI status. As explained in **Section 2.13**, age was not significantly associated with either overall or disease-free survivals in univariable analyses. Yet considering the fact that age is a well-established prognostic marker, especially in overall survival, multivariable models which include age as a covariate are also reported.

Multivariable analysis for polymorphisms associated with overall survival in the univariable analysis

As described in **Section 3.1.1**, the genotypes of the three polymorphisms (*MMP8*-rs12365082, *MMP27*-rs11225388, and *MMP27*-rs11225389) found to be associated with overall survival in the univariable analysis are highly correlated with each other (r_s >0.996). I therefore chose one of these polymorphisms (*MMP27*-rs11225388) to perform the multivariable analysis.

After adjusting for stage and MSI status, patients with an AG or GG genotype of the *MMP27*-rs11225388 polymorphism had lower risk of death than did patients with an AA genotype (**Table 3.7a)**. When adjusted for age at diagnosis, stage, and MSI status

(**Table 3.7b**), a similar result was obtained. As expected, stage and MSI (as well as age) were independent predictors of overall survival.

Results of the multivariable analysis for the haplotypes associated with overall and disease-free survival in the univariable analysis

Of the three haplotypes associated with overall survival and five haplotypes associated with disease-free survival in a univariable analysis, the only association detected in the multivariable analysis was that of the *MMP3* haplotype with overall survival. **Table 3.8** shows the multivariable analysis results for overall survival, and **Appendix H** shows the result for disease-free survival performed for this haplotype. When adjusted only for stage and MSI status, patients with two copies of the most common *MMP3* haplotype had better overall survival (**Table 3.8a**). Also, when adjusted for age at diagnosis, stage and MSI, a similar result was obtained (**Table 3.8b**). **Appendices I-N** show the results for the other haplotypes that were associated with survival in the univariable analyses, but not in the multivariable models.

Table 3.7: Results of the multivariable analysis for the *MMP27* polymorphism and overall survival (dominant genetic model)

a) Adjusting for stage and MSI status (n= 483)

Variable	p-value	HR	95% CI for HR	
			Lower	Upper
Stage	**<0.001**			
Stage II vs stage I	0.116	1.591	0.892	2.84
Stage III vs stage I	**0.003**	2.373	1.345	4.188
Stage IV vs stage 1	**<0.001**	9.398	5.152	17.142
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.189	0.07	0.512
MMP27-rs11225388 AG + GG vs AA	**0.001**	0.581	0.42	0.803

b) Adjusting for stage, age at diagnosis and MSI status (n= 483)

Variable	p-value	HR	95% CI for HR	
			Lower	Upper
Stage	**<0.001**			
Stage II vs stage I	0.118	1.588	0.89	2.834
Stage III vs stage I	**0.002**	2.509	1.419	4.439
Stage IV vs stage 1	**<0.001**	10.417	5.672	19.13
Age at diagnosis	**0.02**	1.021	1.003	1.04
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.191	0.07	0.517
MMP27-rs11225388 AG + GG vs AA	**0.0013**	0.589	0.426	0.814

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable.

Table 3.8: Multivariable analysis for the *MMP3* haplotype associated with overall survival (recessive genetic model)

a) Adjusting for stage and MSI status (n= 483)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.123	1.577	0.883	2.817
Stage III vs I	**0.003**	2.338	1.325	4.127
Stage IV vs I	**<0.001**	9.717	5.335	17.699
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.188	0.069	0.509
**MMP3* haplotype	**0.027**	0.596	0.376	0.943

b) Adjusting for age at diagnosis, stage and MSI status (n= 483)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	**0.016**	1.022	1.004	1.040
Stage	**<0.001**			
Stage II vs I	0.132	1.561	0.874	2.788
Stage III vs I	**0.002**	2.457	1.390	4.343
Stage IV vs I	**<0.001**	10.748	5.866	19.695
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.191	0.070	0.520
**MMP3* haplotype	**0.029**	0.600	0.379	0.950

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable. *patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes.

3.3. Comparison of the entire NFCCR patient cohort (n=736) with patients included in this study (n=505)

In order to determine whether the study cohort (n=505) was representative of the entire NFCCR cohort (n=736), we performed a Chi-square test for categorical variables. This analysis was done for clinicopathological, molecular and treatment-related features including sex, vascular invasion, grade, lymphatic invasion, location, histology, *BRAF* Val600Glu mutation status, MSI status, adjuvant 5-FU based chemotherapy, adjuvant chemotherapy and adjuvant radiation treatment status. The baseline characteristics of the NFCCR cohort are shown in **Appendix O**.

I observed significant differences between the entire NFCCR cohort (n=736) and the patients included in this study (n=505) in terms of the distribution of stage (p-value <0.001). As also shown in **Table 2.1** and **Appendix P**, the study cohort had significantly fewer stage IV patients (9.9%) (**Appendix O**) than the entire NFCCR cohort (20.8%). Significant differences were also detected for lymphatic and vascular invasion status: the entire NFCCR cohort had more patients with vascular invasion or lymphatic invasion when compared to the study cohort (38.3% versus 31.5%, p =0.011 and 38.7% versus 33.1%, p =0.03, respectively) (**Appendices Q** and **R**). We also used the non-parametric Mann-Whitney U-test to compare the median age at diagnosis, and the overall survival and disease free survival times in the two cohorts: the study cohort had longer follow-up times when compared to the entire NFCCR cohort (p <0.001; **Appendix S**).

Chapter 4: Discussion

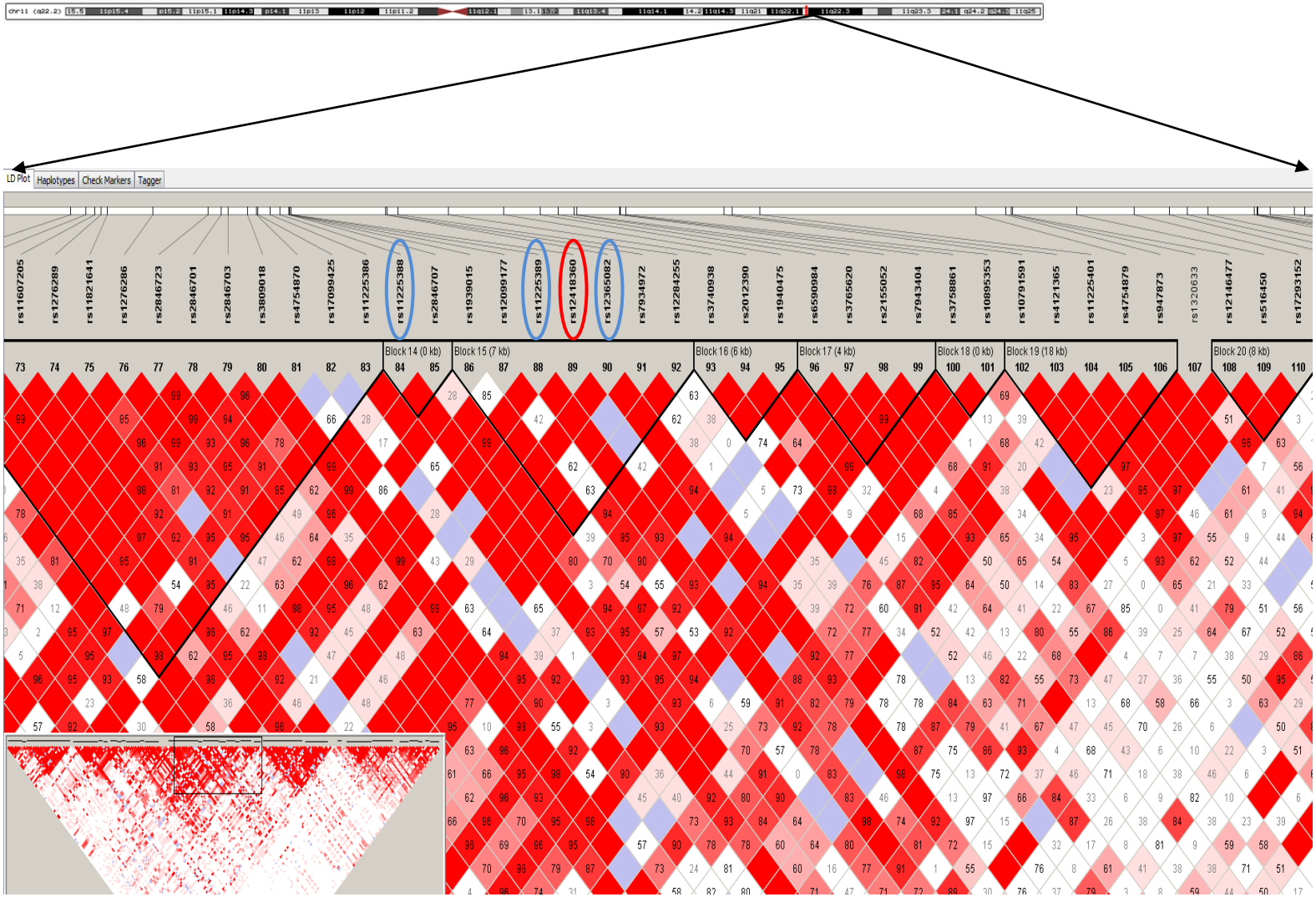
The purpose of this study was to identify new prognostic markers in colorectal cancer. I investigated the associations between survival and 381 genetic polymorphisms (and their combinations in haplotypes) within select genes coding for vascular endothelial growth factors (VEGFs), their receptors (VEGFRs) and matrix metalloproteinases (MMPs).

Substantial biological and clinical data show that variations in angiogenesis, lymph-angiogenesis and metastasis may influence patient survival (9, 10). These processes involve the protein products of several genes, such as members of the VEGF, VEGFR and MMP families. The vascular endothelial growth factor ligands or receptors (e.g. *VEGFA, VEGFR1*) and matrix metalloproteinases (e.g. *MMP1, MMP3,* and *MMP9*) play crucial roles in cancer progression (284, 285) or are associated with survival outcomes in patients (286-290). Due to the established roles of VEGF proteins in carcinogenesis and progression, drugs that target them have been developed for use in patient care (for example Bevacizumab targeting *VEGFA* (291) and Cabozantinib targeting *VEGFR2* (292)). Based on this and other scientific knowledge, this study focused on *VEGF* ligands (n=5), *VEGFR*s (n=3) and all the known human MMP genes (n=23).

The results of the study presented in this thesis suggest that three SNPs *(MMP27*-rs11225389, *MMP27*-rs11225388, and *MMP8*-rs12365082) located within the *MMP8* and *MMP27* genes on chromosome 11q22 are associated with overall survival independent of age at diagnosis, disease stage, and MSI status. These SNPs are potential prognostic

indicators of survival in this disease. Specifically, patients with genotypes containing the minor allele had longer survival time than patients homozygous for the major allele. The genotypes of these three SNPs, which lie within an approximately 7 kb region, are highly correlated with each other (r_s >0.99). The frequency of the minor allele in the study population is about 27%. **Figure 4.1** shows the LD block structure of the genomic region. An intergenic SNP (rs12418360) is located between *MMP8* and *MMP27* (**Figure 4.1**). Since this SNP is intergenic, it was not initially included in this study. A survival analysis was performed for this SNP as well which found no association of this SNP with overall survival under the dominant genetic model (HR =1.362, 95% CI 0.907-2.044, p-value =0.136). Except for one SNP, the genotypes of other SNPs in the two LD regions (LD blocks 14 and 15, **Figure 4.1**) were not highly correlated with the genotypes of these three SNPs (**Appendix T**). While the three SNPs significantly associated with overall survival are almost always inherited together (correlation of their genotypes r_s >0.99), it appears that SNPs not highly correlated with them are not consistently co-inherited. This can happen, for example, if the SNPs represent relatively new mutations. The only SNP correlated with the rs11225388, rs11225389, and rs12365082 SNPs was the *MMP27*-rs2846707 (r_s=0.8, **Appendix T**). However, our analysis did not find it significantly associated with overall survival in a univariable analysis at the pre-specified significance threshold (p-value = 0.0063, HR = 0.658, 95% CI = 0-487-0.889).

a. Chr. 11q22.2


LD Plot
Haplotypes
Check Markers
Tagger
rs11607205
rs1276289
rs11821641
rs1276286
rs2846723
rs2846701
rs2846703
rs3809018
rs4754870
rs17099425
rs11225386
rs11225388
rs2846707
rs1939015
rs12099177
rs11225389
rs12418360
rs12365082
rs7934972
rs12284255
rs3740938
rs2012390
rs1940475
rs6590984
rs3765620
rs2155052
rs7943404
rs3758861
rs10895353
rs10791591
rs4121365
rs11225401
rs4754879
rs947873
rs1320633
rs12146477
rs516450
rs17293152
Block 14 (0 kb)
Block 15 (7 kb)
Block 16 (6 kb)
Block 17 (4 kb)
Block 18 (0 kb)
Block 19 (18 kb)
Block 20 (8 kb)

b. A close view of LD blocks 14 and 15 showing the four SNPs.

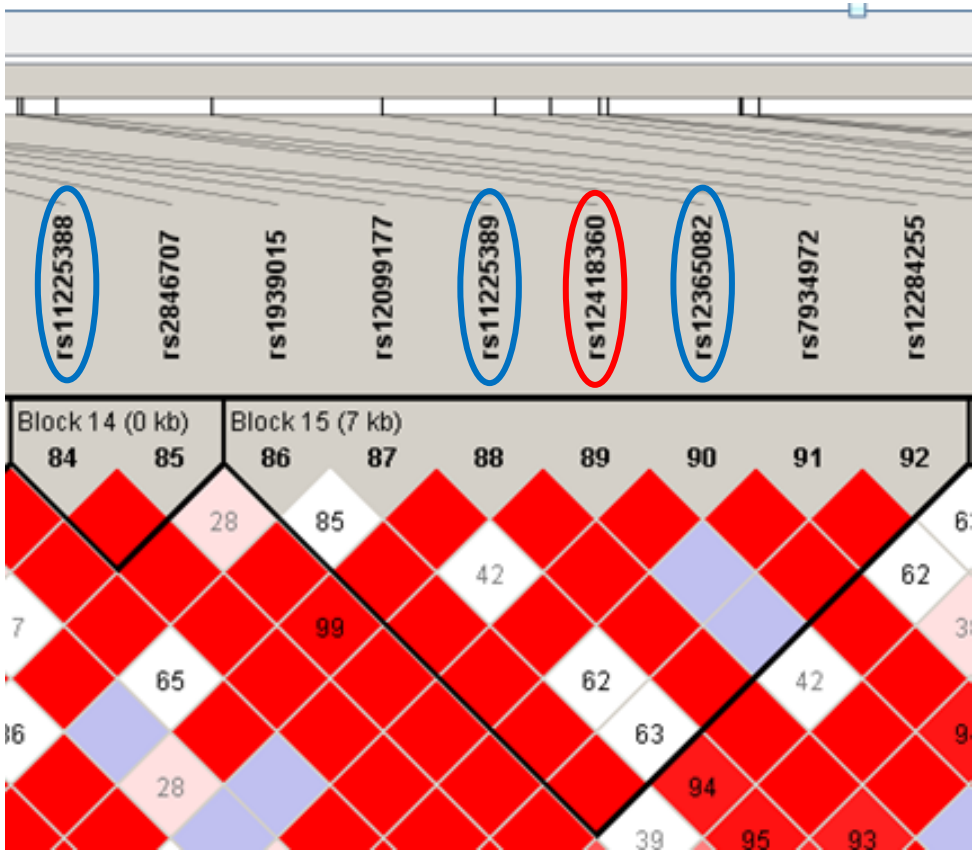


Figure 4.1a: LD block structure of the genomic region containing the nine MMP genes (*MMP7*, *MMP20*, *MMP27*, *MMP8*, *MMP10*, *MMP1*, *MMP3*, *MMP12*, and *MMP13*) on chromosome 11q22. **Figure 4.1b**: The blue circles show the three SNPs (rs11225389, rs11225388, and rs12365082) that were found to be associated with overall survival in this study (in LD blocks 14 and 15). The red circle indicates the intergenic SNP (rs12418360).

The *MMP27*-rs11225389, *MMP27*-rs11225388, and *MMP8*-rs12365082 polymorphisms are all located in non-coding regions, specifically the 5'-UTR (*MMP27*-rs11225389), 3'-UTR (*MMP8*-rs12365082) and intronic regions (*MMP27*-rs11225388). As of July 2014, there is no published report concerning their potential biological significance. According to a computational tool, snpinfor (293), these SNPs are predicted to be located within biologically functional regions. For example, the *MMP27*-rs11225388 and *MMP27*-rs11225389 polymorphisms are located in binding sites of several transcription factors such as AP1, CDPCR3, TAXCREB and AP4, PAX6, PPARG, respectively (293). The *MMP8*-rs12365082 polymorphism is located in a binding site of a miRNA, hsa-miR (293). Thus, these polymorphisms may affect the expression levels of these genes. Further studies are needed to test the biological roles of these SNPs and their relation to progression in colorectal cancer. In addition, there is no previous report addressing the associations of these particular SNPs with clinical outcomes in colorectal cancer. To our knowledge, this is the first time these polymorphisms have been investigated and found to be associated with outcome in colorectal cancer.

Four of the polymorphisms included in this study have previously been studied in relation to survival outcomes in colorectal cancer. Dassoulas *et al*. (227) found an association between the (*VEGFA* +936 C/T (rs3025039; NM_001025366.1:c.*237C>T) polymorphism) and overall survival. However, this association was not confirmed in the present study or in other studies (231, 233 and 265). In the case of the *VEGFA*-634 G/C (rs2010963; NM_001025366.1:c.-94C>G) polymorphism, the same group reported its association with overall survival (227). However, neither our study nor other studies

(231-233, 265) replicated this finding. These conflicting results between the present and other study may be due to the differences in patient ethnicities, treatment characteristics of the cohorts, the study design or the statistical approaches used, (such as the p-value threshold that defined the significance level). In addition, our present study found no association between another polymorphism (*KDR* 1192 C/T (rs2305948; NP_002244.1:p.Val297Ile)) with survival times as previously reported (232). Another graduate student in our laboratory had previously investigated the associations of *VEGFA*-634 G/C (rs2010963) and *VEGFA* +936 C/T (rs3025039; NM_001025366.2:c.*237C>T) SNPs, also included in this study, in a similar NFCCR patient sub-cohort (265). That study analyzed the genotypes using a co-dominant genetic model but, similar to our results, found no association of these SNPs with clinical outcome.

To complement the single-SNP survival association approach, I also performed gene-based haplotype analysis, using phased haplotypes for each patient. The result of this analysis showed that one haplotype (in the *MMP3* gene) was significantly associated with overall survival in patients when adjusted for other prognostic variables. This *MMP3* haplotype contains four SNPs: rs566125, rs3025066, rs3020919, and rs679620. To our knowledge, this is the first study to investigate and identify this haplotype as associated with outcome in colorectal cancer. The biological relevance of this haplotype to the risk of death in colorectal cancer patients is yet to be established.

So far, very few studies have tested the associations of haplotypes with survival outcomes in colorectal cancer. Kim *et al.* (15) showed that a *VEGFA* haplotype consisting of the -2578C/A (rs699947; NM_001025366.2:c.-2055A>C), -634G/C (rs2010963;

NM_001025366.1:c.-94C>G), and +936C/T (rs3025039; NM_001025366.2:c.*237C>T) polymorphisms was associated with outcome in colorectal cancer patients. Hansen *et al*. (14) showed that a haplotype consisting of the *VEGFA* -2578C/A (rs699947; NM_001025366.2:c.-2055A>C), -460C/T (rs833061; NM_001025366.2:c.-958C>T) and 405G/C (rs2010963; NM_001025366.2:c.-94C>G) polymorphisms was significantly associated with survival in a cohort of colorectal cancer patients. These studies may not be directly comparable to the present study which used different sets of SNPs and haplotypes.

Interestingly, both the single SNP and the haplotype analysis in this study identified associations between the three matrix metalloproteinase genes (*MMP8*, *MMP27*, and *MMP3*) and overall survival in colorectal cancer. MMP8 also called neutrophil collagenase is mainly expressed in neutrophils. The MMP8 protein belongs to a group of extracellular proteases that have the ability to degrade the extracellular matrix (294). The role of MMP8 is the degradation of type I, II and III collagens. The second gene identified in this study, *MMP27*, encodes a matrix metalloproteinase that helps degrade extracellular matrix components such as fibronectin, gelatins and aggrecan (295). Somatic mutations of the *MMP8* and *MMP27* genes have been reported in some cancers (e.g. thyroid cancer, or melanoma) but not previously in colorectal cancer (296-298). *MMP3* is another matrix metalloproteinase gene associated with survival outcomes in this study. The MMP3 protein degrades components of the extracellular matrix such as collagen IV, fibronectin, proteoglycan, and laminin (299). Many reports have associated mutations of this gene with diseases such as colorectal cancer (300), myocardial infarction (301), Takayasu arteritis (302), Alzheimer's disease (303), and gastric cancer

(304). Interestingly, *MMP3*, *MMP8*, and *MMP27* are all located in a MMP gene cluster on chromosome 11q22 (305) (**Figure 4.1a**). This is the first report that suggests an association between this chromosomal region and the risk of death in colorectal cancer.

I am aware of the limitations of this study. Since we considered only common genetic variants and haplotypes (frequencies ≥ 5%) in the study population, I may have missed rare genetic variations or haplotypes that could have strong effects on prognosis. Similarly, in the haplotype analysis, I tested only the associations of the most common haplotype for each gene compared to other haplotypes. The potential prognostic associations of other individual haplotypes remain to be tested. Our study cohort is biased towards early stage patients. Stage IV patients and those with vascular or lymphatic invasion of the tumor are underrepresented when compared to the entire NFCCR cohort. This is because many late-stage patients were already deceased before being enrolled into the study and therefore no blood sample could be obtained for DNA extraction (deceased patients could be enrolled into the NFCCR cohort by proxy consent from a relative). It is also not clear why I did not identify age as a prognostic factor in the univariate analysis, but it can hypothesized that the bias described for the study cohort may have a role in it. I did not analyze all the SNPs in these gene regions, either because they were not present on the Illumina SNP genotyping platform or because they failed to be genotyped in the patient cohort. Fifth, not all the genes functioning in the angiogenesis, lymph-angiogenesis or metastasis pathways were investigated in relation to outcome. Sixth, the patient cohort consisted of Caucasian patients only, thus the results may not be relevant to colorectal cancer patients from other human populations. I am also aware that the *MMP8* and *MMP27* SNPs, as well as of the *MMP3* haplotype, found to be associated with

survival in the study cohort may be false-positive associations. Thus, one of the future research aims of our laboratory is to replicate these associations in an additional patient cohort previously collected between 1997-1998 in Newfoundland.

This study also has many strengths. First, I investigated a relatively large number of patients compared to the majority of outcome studies previously published. Second, the follow-up period was relatively long, allowing us to accumulate a large number of events of interest (i.e. occurrence of death, recurrence and metastasis). Third, stringent quality control procedures were implemented to limit potentially erroneous genotype data and patient mix-up. Fourth, this is the first study that comprehensively examined a large number of polymorphisms within multiple VEGF ligand, VEGF receptor, and matrix metalloproteinase genes in relation to outcome in colorectal cancer. Fifth, to my knowledge, our laboratory is the first, if not the only, laboratory in Canada that investigates genetic polymorphisms as candidate prognostic markers in colorectal cancer (265, 306 and 307).

Conclusion

In a cohort of colorectal cancer patients from Newfoundland, I conducted a candidate-pathway survival association study involving 381 polymorphisms within 30 key angiogenesis, lymph-angiogenesis and metastasis genes. Three highly correlated SNPs (*MMP27*-rs11225388 G/A, *MMP27*-rs112253389 A/C, and *MMP8*-rs12365082 T/A) located in two MMP genes (*MMP8* and *MMP27*) were found to be associated with

overall survival independent of other prognostic markers. Analyzing the combined effects of SNPs in the form of haplotypes with patient outcome, I was also able to find an association with overall survival and a *MMP3* haplotype. The biological relevance of these three SNP and the *MMP3* haplotype to the risk of death remains to be established. Future studies are needed to validate these associations and to ascertain the biological mechanisms underlying the effects of these polymorphisms and haplotypes on survival of colorectal cancer patients.

References

1. World Cancer Research Fund and American Institute for Cancer Research Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective (2007). Washington, DC: American Institute for Cancer Research. http://www.dietandcancerreport.org/cancer_resource_center/downloads/Second_Expert_Report_full.pdf [Accessed April 23rd, 2013].
2. Boyle, P., Langman, J.S. (2000). ABC of colorectal cancer: Epidemiology. *BMJ* 321, 805-808.
3. Colorectal Cancer Statistics (2013). Featuring colorectal cancer Canadian cancer Society's Steering Committee on Cancer Statistics, Toronto, ON, CAN. http://www.cancer.ca/~/media/cancer.ca/CW/cancer%20information/cancer%20101/Canadian%20cancer%20statistics/canadian-cancer-statistics-2013-EN.pdf [Accessed April 23rd, 2014].
4. Gryfe, R. (2009). Inherited colorectal cancer syndromes. *Clin Colon Rectal Surg* 22, 198-208.
5. American Cancer Society. Colorectal cancer facts and figures 2011-2013. Atlanta: American Cancer Society, 2011. [Accessed June 14th, 2013]. http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-028323.pdf
6. Compton, C.C., Fielding, L.P., Burgart, L.J. *et al*. (2000). Prognostic factors in colorectal cancer: College of American Pathologists Consensus Satement 1999. *Arch Pathol Lab Med* 124, 979-994.
7. Folkman, J. (2000). Tumor angiogenesis. In: Holland, J.F., *et al.* (Ed.), Cancer Medicine. B. C. Decker Inc., Ontario, Canada, pp 132–152.
8. Albrecht, I., Christofori, G. (2014). Molecular mechanisms of lymph-angiogenesis in development and cancer. *Int J Dev Biol* 55, 483-494.
9. Plate, K.H. (2001). From angiogenesis to lymph-angiogenesis. *Nat Med* 7, 151-152.

10. Alitalo, A., Detmar, M. (2012). Interaction of tumor cells and lymphatic vessels in cancer progression. *Oncogene* 31, 4499-4508.
11. Cargill, M., Altshuler, D., Ireland, J. *et al*. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22, 231-238.
12. Melzer, D. (2008). Genetic polymorphisms and human aging: association studies deliver. *Rejuvenation Res* 11, 523–526.
13. Shastry, B.S. (2007). SNPs in disease gene mapping, medical drug development and evolution. *J Hum Genet* 52, 871–880.
14. Hansen, T.F., Spindler, K.L.G., Andersen, R.F. *et al*. (2010). The prognostic value of haplotypes in the vascular endothelial growth factor A gene in colorectal cancer. *Cancer (Basel)* 2, 1405-1418.
15. Kim, J.G., Chae, Y.S., Sohn, S.K. *et al*. (2008). Vascular endothelia growth factor gene polymorphisms associated with prognosis for patients with colorectal cancer. *Clin Cancer Res* 14, 62-66.
16. Colorectal cancer (2012). Enyclopedia Britannica Online. Accessed 07 July, 2012 from: http://www.britannica.com/EBchecked/topic/126587/colorectal-cancer.
17. Spann, S., Levin, B., Rozen, P. *et al*: Colorectal cancer: How big is the problem, why prevent it and how might it present? In Rozen P., Young, G., Levin, B. *et al* (eds): Colorectal Cancer in Clinical Practice: Prevention, Early Detection and Mangement. London, Martin Dunitz Ltd, 2002, pp 1-13.
18. Hanahan, D., Weinberg, R.A. (2000). The hallmarkers of cancer. *Cell* 100, 57-70.
19. Ellis, C., Saddler, D: Colorectal cancer, in Yarbro, C.H., Frogge, M.H., Goodman, M., Groenwald, S (eds): Cancer Nursing: Principles and Practice (ed 5). Boston, M.A, Jones and Bartlett, 2000, pp 1117-1137.
20. Jemal, A., Clegg, L.X., Ward, E. *et al*. (2004). Annual report to the nation on the status of cancer, 1975-2001, with a special feature regarding survival. *Cancer* 101, 3–27.

21. Ries, L.A.G., Melbert, D., Krapcho, M. *et al*. (2008). SEER cancer statistics review, 1975–2005, National Cancer Institute. Bethesda. MD, http://seer.cancer.gov/csr/1975_2005/ [Accessed June 20, 2012].
22. Bernold, D.M., Sinicrope, F.A. (2006). Advances in Chemotherapy for Colorectal Cancer. *Clin Gastroenterol Hepatol* 4, 808-821.
23. Hawk, E. T., Levin, B. (2005). "Colorectal cancer prevention". *J Clin Oncol* 23, 378–391.
24. Kinzler, K.W., Vogelstein, B. (2002). Colorectal tumors. In: Vogelstein B, Kinzler KW, editors. The genetic basis of human cancer. 2nd ed McGraw-Hill; *New York:* pp. 583–612.
25. Markowitz, S.D., Dawson, D.M., Willis, J. *et al*. (2002). Focus on colon cancer. *Cancer Cell* 1, 233–236.
26. Flood, D.M., Weiss, N.S., Cook, L.S. *et al*. (2000). Colorectal cancer incidence in Asian migrants to the United States and their descendants. *Cancer Causes Control* 11, 403-411.
27. Norat, L., Bingham, S., Ferrari, P. *et al*. (2005). Meat, fish, and colorectal cancer risk: the European prospective investigation into cancer and nutrition. *J Natl Cancer Inst* 97, 906-916.
28. Bird, C., Swendseid, M., Witte, J. *et al*. (1995). Red cell and plasma folate, folate consumption and risk of colorectal adenomatous polps. *Cancer Epidemiol Biomarkers Prev* 4, 709-714.
29. Paspatis, G.A., Kalafatis, E., Oros, L. *et al*. (1995). Folate status and adenomatous colonic polyps. A colonoscopically controlled study. *Dis Colon Rectum* 38, 64-67.
30. Baron, J.A., Sandler, R.S., Haile, R.W. *et al.* (1998). Folate intake, alcohol consumption, cigarette smoking, and risk of colorectal adenomas. *J Natl Cancer Inst* 90, 57-62.
31. National Institutes of Health What You Need To Know About Cancer of the Colon and Rectum. Bethesda, MD: U.S. Department of Health and Human Services and National Institutes of Health; 2006. [Accessed July 20, 2012] http://www.cinj.org/sites/cinj/files/documents/WYNTK_Colon.pdf

32. Gingras, D., Belivea, R. (2011). Colorectal cancer prevention through dietary and lifestyle modifications. *Cancer Microenviron* 4, 133-139.
33. Lee, K.J., Inoue, M., Otani, T. *et al*. (2007). Physical activity and risk of colorectal cancer in Japanese men and women: the Japan Public Health Center-based prospective study. *Cancer causes Control* 18, 199-209.
34. Botteri, E., Iodice, S., Raimondi, S. *et al.* (2008). Cigarette smoking and adenomatous polyps: a meta-analysis. *Gastroenterology* 134, 388–395.
35. Pöschl G, Seitz H K. (2004). Alcohol and cancer. *Alcohol* and *Alcoholism* 39, 155–165.
36. Lynch, H.T., de la Chapelle, A. (2003). Hereditary colorectal cancer. *N Engl J Med* 348, 919-932.
37. Aaltonen, A., Salovaara, R., Kristo, P. *et al.* (1998). Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N Engl J Med* 338, 1481–1487.
38. Houlston, R.S., Web, E., Broderick, P. *et al*. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat Genet* 40, 1426-1435.
39. De la Chapelle (2004). Genetic predisposition to colorectal cancer. *Nat Rev Cancer* 4, 769–780.
40. Stigliano, V., Assisi, D., Cosimelli, M. *et al*. (2008). Survival of hereditary non-polyposis colorectal cancer patients compared with sporadic colorectal cancer patients. *J Experim Clin Res* 27:39.
41. Naccarati, A., Pardini, B., Hemminiki, K. *et al*. (2007). 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.
42. Kury, S., Buecher, B., Robiou-du-Pont, S. *et al*. (2008). Low-penetrance alleles predisposing to sporadic colorectal cancers: a French case-controlled genetic association study. *BMC* 8:326.

43. Gai, P., Maryan, N., Hennig, E.E. *et al*. (2012). Pooled sample-based GWAS: a cost- effective alternative for identifying colorectal and prostate cancer risk variants in Polish population. *Plos One* 7(4): e35307.
44. Dunlop, M.G., Dobbins, S.E., Farrington, E. *et al*. (2012). Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet* 44, 770-776.
45. Campbell, W.J., Spence, R.A., Parks, T.G. (1994). Familial adenomatous polyposis. Br J Surg 81, 1722-1733.
46. Bussey, H.J., Veale, A.M., Morson, B.C. (1978). Genetics of gastrointestinal polyposis. *Gastroenterology* 74, 1325–1330.
47. Groden, J., Thliveris, A., Samowitz, W. *et al*. (1991). Identification and characterization of the familial adenomatous polyposis coli gene. *Cell* 66, 589–600.
48. Kinzler, K.W., Nilbert, M.C., Su, L.K. *et al*. (1991). Identification of FAP locus genes from chromosome 5q21. *Science* 253, 661–665.
49. Chung, D.C. (2000). The genetic basis of colorectal cancer: insights into critical pathways of tumorigenesis. *Gastroenterology* 119, 854–865.
50. Eccles, D.M., Lunt, P.W., Wallis, Y. *et al*. (1997). An unusually severe phenotype for familial adenomatous polyposis. *Arch Dis Childhood* 77, 431-435.
51. Jasperson, K.W., Burt, R.W. APC-Associated Polyposis Conditions. 1998 Dec 18[Updated 2014 Mar 27]. In: Pagon RA, Adam MP, Ardinger HH, *et al*., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2014. Available from: http://www.ncbi.nlm.nih.gov/books/NBK1345/
52. Bjork, J. Akerbrant, H. Iselius, L. *et al*. (2001). Periampullary adenomas and adenocarcinomas in familial adenomatous polyposis: cumulative risks and APC gene mutations. *Gastroenterology* 121, 1127–1135.
53. Groves, C.J., Saunders, B.P., Spigelman, A. D. *et al*. (2002). Duodenal cancer in patients with familial adenomatous polyposis (FAP): results of a 10 year prospective study. *Gut* 50, 636–641.

54. Friedl, W., Meushel, S., Caspari, R. *et al*. (1996). Attenuated familial adenomatous polyosis due to a mutation in the 3' part of the APC gene. A clue for understanding the function of the APC protein. *Hum Genet* 97, 579-584.
55. Spirio, L., Otterud, B., Stauffer. D. *et al*. (1992). Linkage of a variant or attenuated form of adenomatous polyposis coli to the adenomatous polyposis coli (APC) locus. *Am J Hum Genet* 51, 92-100.
56. Juhn, E., Khachemoune, A. (2010). Gardner syndrome: skin manifestations, differential diagnosis and management. *Am J Clin Dermatol* 11, 117–122.
57. Lebrun, C., Olschwang, S., Jeannin, S. *et al*. (2007). Turcot syndrome confirmed with molecular analysis. *Eur J Neurol* 14, 470–472.
58. Mori, T., Nagase, H., Horii, A. *et al*. (1994). Germ-line and somatic mutations of the APC gene in patients with Turcot syndrome and analysis of APC mutations in brain tumors. *Genes Chromosomes Cancer* 9, 168-172.
59. Nielsen, M., Morreau, H., Vasen, H.F. *et al*. (2010). *MUTYH*-associated polyposis (MAP). *Crit Rev Oncol Hematol* 79, 1-16.
60. Croitoru, M.E., Cleary, S.P., Di Nicola, N. *et al*. (2004). Association between biallelic and monoallelic germline MYH gene mutations and colorectal cancer risk. *J Natl Cancer Inst* 96, 1631-1634.
61. Lubbe, S.J., Di Bernardo, M.C., Chandler, I.P. *et al*. (2009). Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol* 27, 3975–3980.
62. Jenkins, M.A., Croitoru, M.E., Monga, N. *et al*. (2006). Risk of colorectal cancer in monoallelic and biallelic carriers of MYH mutations: a population-based case-family study. *Cancer Epidemiol Biomarkers Prev* 15, 312–314.
63. Slupska, M.M., Baikalov, C., Luther, W.M. *et al*. (1996). Cloning and sequencing a human homolog (hMYH) of the Escherichia coli mutY gene whose function is required for the repair of oxidative DNA damage. *J Bacteriol* 178, 3885–3892.
64. Al-Tassan, N., Chmiel, N.H., Maynard, J. *et al*. (2002). Inherited variants of MYH associated with somatic G: C→T: A mutations in colorectal tumors. *Nat Genet* 30, 227–232.

65. Rubio, C.A., Stemme, S., Jaramillo, E. *et al*. (2006). Hyperplastic polyposis coli syndrome and colorectal carcinoma. *Endoscopy* 38, 266–270.
66. Chow, E., Lipton, L., Lynch, E., *et al*. (2006). Hyperplastic polyposis syndrome: phenotypic presentations and the role of MBD4 and MYH. *Gastroenterology* 131, 30-39.
67. Young, J., Jenkins, M., Parry, S. *et al*. (2007). Serrated pathway colorectal cancer in the population: genetic consideration. *Gut* 56, 1453–1459.
68. Coburn, M.C., Pricolo, V.E., DeLuca, F.G. *et al*. (1995). Malignant potential in intestinal Juvenile polyposis syndromes. *Ann Surg Oncol* 2, 386–391.
69. Jass, J.R., Williams, C.B., Bussey, H.J. *et al*. (1988). Juvenile polyposis: a precancerous condition. *Histopathol* 13, 619-630.
70. Watanabe, A., Magashima, H., Motoi, M. *et al*. (1979). Familial juvenile polyposis of the stomach. *Gastroenterol* 77, 148-151.
71. Howe, J.R., Mitros, F.A., Summers, R.W (1998). The risk of gastrointestinal carcinoma in familial juvenile polyposis. *Ann Surg Oncol* 5, 751-756.
72. Howe, J.R., Roth, S., Ringold, J.C. *et al*. (1998). Mutations in the SMAD4/DPC4 gene in juvenile polyposis. *Science* 280, 1086–1088.
73. Howe, J.R., Bair, J.L, Sayed, M.G. *et al*. (2001). Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. *Nat Genet* 28, 184–187.
74. Liaw, D., Marsh, D.J., Li, J. *et al*. (1997). Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome. *Nat Genet* 16, 64–67.
75. Farooq, A., Walker, L.J., Bowling, J. *et al*. (2010). Cowden syndrome. *Cancer Treat Rev* 36, 577–583.
76. Kopacova, M., Tacheci, I., Rejchrt, S. *et al*. (2009). Peutz-Jeghers syndrome: diagnostic and therapeutic approach. *World J Gastroenterol* 15, 5397–5408.
77. Giardiello, F.M., Brensinger, J.D., Tersmette, A.C. *et al*. (2000). Very high risk of cancer in familial Peutz-Jeghers syndrome. *Gastroenterology* 119, 1447–1453.

78. Abdel-Rahman W.M., Peltomäki, P. (2008). Lynch syndrome and related familial colorectal cancers. *Crit Rev Oncog* 14, 1–22.
79. Samowitz, W.S., Curtin, K., Lin, H.H. *et al*. (2001). The colon cancer burden of genetically defined hereditary nonpolyposis colon cancer. *Gastroenterology* 121, 830–838.
80. Fishel, R., Lescoe, M.K., Rao, M.R. *et al*. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* 75, 1027–1038.
81. Bronner, C.E., Baker, S.M., Morrison, P.T. *et al*. (1994). Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* 368, 258–261.
82. Leach, F.S., Nicolaides, N.C., Papadopoulos, N. *et al*. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* 75, 1215–1225.
83. Papadopoulos, N., Nicolaides, N.C., Wei, Y.F. *et al*. (1994). Mutation of a mutL homolog in hereditary colon cancer. *Science* 263, 1625–1629.
84. Nicolaides, N.C., Papadopoulos, N., Liu, B. *et al*. (1994). Mutations of two PMS homologues in hereditary non-polyposis colon cancer. *Nature* 371, 75–80.
85. Liu, B., Parsons, R.E., Hamilton, S.R. *et al*. (1994). hMSH2 mutations in hereditary nonpolyposis colorectal cancer kindreds. *Cancer Res* 54, 4590–4594.
86. Peltomaki, P., Aaltonen, L.A., Sistonen, P. *et al.* (1993). Genetic mapping of a locus predisposing to human colorectal cancer. *Science* 260, 810–812.
87. Lindblom, A., Tannergard, P., Werelius, B. *et al*. (1993). Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nat Genet* 5, 279–282.
88. Prolla, T.A., Pang, Q., Alani, E. *et al*. (1994). MLH1, PMS1, and MSH2 interactions during the initiation of DNA mismatch repair in yeast. *Science* 265, 1091–1093.
89. Burt, R.W., Barthel, J.S., Dunn, K.B. *et al*. (2010). National Comprehensive Cancer Network: Clinical practice guidelines in oncology- colorectal cancer screening. *J Natl Compr Can Netw* 8, 8-61.

90. Lindor, N.M., Rabe, K., Petersen, G.M. *et al*. (2005). Lower Cancer Incidence in amsterdam-I criteria families without mismatch repair deficiency: Familial colorectal cancer type X. *JAMA* 293, 1979-1985.
91. Compton, C.C. (2000). Updated protocol for the examination of specimens from patients with carcinomas of the colon and the rectum, excluding carcinoid tumors, lymphomas, sarcoma, and tumors of the vermiform appendix: a basis for checklists. *Cancer Committee Arch Pthol Lab Med,* 124, 1016.
92. O'Connell, J.B., Maggard, M.A., Ko, C.Y. (2004). Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *J Natl Cancer Inst* 96, 1420-1425.
93. Wiggers, T., Arends, J.W., Volovics, A. (1998). Regression analysis of prognostic factors in colorectal cancer after curative resection. *Dis Colon Retum* 31, 33-41.
94. American Cancer Society. Colorectal cancer facts and figures 2011-2013. Atlanta: American Cancer Society, 2011. [Accessed June 14th, 2013]. http://www.cancer.org/acs/groups/content/@epidemiologysurveilance/documents/document/acspc-028323.pdf
95. Compton, C.C., Greene, F.L. (2004). The staging of colorectal cancer: 2004 and beyond. *CA Cancer J Clin* 54, 295-308.
96. Greene, F.L., Page, D.L., Fleming, I.D. *et al*. AJCC Cancer Staging Manual. 6th Ed. New York, NY: Springer; 2002.
97. Sobin, L.H., Witttekind, C. (eds.). TNM: Classification of maglignant tumours. 6th Ed. New York, NY: Wiley-Liss; 2002.
98. Swanson, R.S., Compton, C.C., Stewart, A.K. *et al*. (2003). The prognosis of T3N0 colon cancer is dependent on the number of lymph nodes examined. *Ann Surg Oncol* 10, 65-71.
99. Chen, S.L., Bilchik, A.J. (2006). More extensive nodal dissection improves survival for stages I to III of colon cancer: a population-based study. *Ann Surg* 244, 602-610.
100. De Campos-Lobato, L.F., Stocchi, L., de Sousa, J.B. *et al*. (2013). Less than 12 nodes in the surgical specimen after total mesorectal excision following

neoadjuvant chemoradiation: it means more than you think. *Ann Surg Oncol* 20, 3398-3406.

101. Betge, J., Pollheimer, M.J., Lindtner, R.A. *et al*. (2012). Intramural and extramural vascular invasion in colorectal cancer prognostic significance and quality of pathology reporting. *Cancer* 118, 628-638.
102. Bayar, S., Saxena, R., Emir, B. *et al*. (2002). Venous invasion may predict lymph node metastasis in early rectal cancer. *Eur J Surg Oncol* 28, 413-417.
103. Sato, T., Ueno, H., Mochizuki, H., *et al*. (2010). Objective criteria for the grading of venous invasion in colorectal cancer. *Am J Surg Pathol* 34, 454-462.
104. Suzuki, A., Togashi, K., Nokubi, M., *et al*. (2009). Evaluation of venous invasion by Elastica van Gieson stain and tumor budding predicts local and distant metastases in patients with T1 stage colorectal cancer. *Am J Surg Pathol* 33, 1601-1607.
105. Ouchi, K., Sugawara, T., Ono, H., *et al*. (1996). Histologic features and clinical significance of venous invasion in colorectal carcinoma with hepatic metastasis. *Cancer* 78, 2313-2317.
106. Rahbari, N.N., Bork, U., Motschall, E. *et al*. (2012). Molecular detection of tumor cells in regional lymph nodes is associated with disease recurrence and poor survival in node- negative colorectal cancer: a systematic review and meta-analysis. *J Clin Oncol* 30, 60-70.
107. Mulcahy, H.E., Skelly, M.M., Husain, A. *et al*. (1996). Long-term outcome following curative surgery for malignant large bowel obstruction. *Br J Surg* 83, 46-50.
108. Minsky, B.D., Miles, C., Recht, A. (1989). Lymphatic vessel invasion is an independent prognostic factor for survival in colorectal cancer. *Int J Radiat Oncol Biol Phys* 17, 311-318.
109. Internation Union Against Cancer (UICC): TNM Classification of Malignant Tumor 4th ed. Hermaek P, Sobin, L H, eds. Heidelberg: Springer; 1987. Revised. 1992.

110. American Joint Committee on Cancer: Manual for staging on cancer, 4th ed. J.B. Lippincott Company, Philadelphia 1993, pp 161-168.
111. Hyngstrom, J.R., Hu, C.Y., Xing, Y. *et al*. (2012). Clinicopathology and outcomes for mucinous and signet ring colorectal adenocarcinoma: analysis from the National Cancer Database. *Ann Surg Oncol* 19, 2814-2821.
112. Griffin, M.R., Bergstralh, E.J., Coffey, R.J. *et al*. (1987). Predictors of survival after curative resection of carcinoma of the colon and rectum. *Cancer* 60, 2318-2324.
113. Scott, N.A., Wie, H.S., Moertel, C.G. *et al.* (1987). Colorectal cancer. Dukes' stage, tumor Site, preoperative plasma CEA level, and patient prognosis related to tumor DNA ploidy pattern. *Arch Surg* 122, 1375-1379.
114. Fisher, E.R., Sass, R., Palekar, A. *et al.* (1989). Dukes classification revisited. Findings from the National Surgical Adjuvant Breast and Bowel Projects (Protocol R-01). *Cancer* 64, 2354-2360.
115. Pawlik, T.M., Raut, C.P., Rodriguez-Bigas, M.A. (2004). Colorectal carcinogenesis: MSI-H versus MSI-L. *Dis Markers* 20, 199-206.
116. Pallai, R. (1992). Oncogenes and oncoproteins as tumor markers. *Eur J Surg Oncol* 18, 417-424.
117. Der, C.J., Krontiris, T.A., Cooper, G.M. (1982). Transforming genes of human bladder and lung carcinoma cell lines are homologous to the *ras* genes of Harvey and Kristen sarcoma viruses. *Proc Natl Acad Sci USA* 79, 3637-3640.
118. Shimizu, K., Goldfarb, M., Perucho, M. *et al*. (1983). Isolation and preliminary characterization of the transforming gene of a human neuroblastoma cell line. *Proc Natl Acad Sci USA* 80, 383-387.
119. Vogelstein, B, Fearon, E.R., Hamilton, S.R. *et al*. (1988). Genetic alterations during colorectal-tumor development. *N Engl J Med* 319, 525-532.
120. Barbacid, M. (1987). Ras genes. *Annu Rev Biochem* 56, 779-827.
121. Benhatter, J., Losi, L., Chaubert, P. *et al*. (1993). Prognostic significance of K-ras mutations in colorectal carcinoma. *Gastroenterology* 104, 1044-1048.

122. Markowitz, S., Hines, J.D., Lutterbaugh, J. *et al*. (1995). Mutant K-ras oncogenes in colon cancers do not predict patients' chemotherapy response or survival. *Clin Cancer Res* 1, 441-445.
123. Dix, B.R., Robbins, P., Soong, R. *et al*. (1994). The common molecular genetic alterations in Dukes' B and C colorectal carcinomas are not short-term prognostic indicators of survival. *Int J Cancer* 59, 747-751.
124. McBride, O.W., Merry, Givol D. (1986). The gene for human p53 cellular tumor antigen is located on chromosome 17 short arm (17p13). *Proc Natl Acad Sci USA* 83, 130-134.
125. Lane, D.P. (1992). Cancer. P53 guardian of the genome. *Nature* 358, 15-16.
126. Nigro, J.M., Baker, S.J., Preisinger, A.C. *et al*. (1989). Mutations in the p53 gene occur in diverse human tumour types. *Nature* 342, 705-708.
127. Hao, X.P., Frayling, I.M., Sgouros, L.G. *et al*. (2002). The spectrum of p53 mutations in colorectal adenomas from that in colorectal carcinomas. *Gut* 50, 834-839.
128. Tortola, S., Marcuello, E., Gonzalez, I. *et al*. (1999). P53 and K-ras gene mutations correlate with tumor aggressiveness but are not of routine prognostic value in colorectal cancer. *J Clin Oncol* 17, 1375-1381.
129. Yamaguchi, A., Kurosaka, Y., Fushida, S. *et al*. (1992). Expression of p53 protein in colorectal cancer and its relationship to short term prognosis. *Cancer* 70, 2778-2784.
130. Soong, R., Powell, B., Elsaleh, H. *et al.* (2000). Prognostic significance of *TP53* gene mutation in 995 cases of colorectal Carcinoma: Influence of tumour site, stage, adjuvant chemotherapy and type of mutation. *Eur J Cancer* 36, 2053–2060.
131. Elsaleh, H., Powell, B., Soontrapornchai, P. *et al*. (2000). *p53* gene mutation microsatellite instability and adjuvant chemotherapy: impact on survival of 388 pateints with Dukes' C colon carcinoma. *Oncology* 58, 52–59.
132. Bowman, B.M., Wildrick, D.M., Alfaro, S.R. (1988). Chromosome 18 allele loss at the D18S6 lous in human colorectal carcinomas. *Biochem Biophys Res Commun* 155, 463-469.

133. Popat, S., Houlston, R.S. (2005). A systematic review and meta-analysis of the relationship between chromosome 18q genotype, DCC status and colorectal cancer prognosis. *Eur J Cancer* 41, 2060-2070.
134. Jernvall, P., Makinen, M.J., Karttunen, T.J. *et al*. (1999). Loss of heterozygosity at 18q21 is indicative of recurrence and therefore poor prognosis in a subset of colorectal cancers. *Br J Cancer* 79, 903–908.
135. Carethers, J.M., Hawn, M.T., Greenson. J.K. *et al*. (1998). Prognostic significance of allelic loss at chromosome 18q21 for stage II colorectal cancer. *Gastroenterology* 114, 1188–1195.
136. The NME4 gene. http://genes.mit.edu/cgi-bin/targetscan_lookup2.pl?KEYWORD=let-7/miR-98 [Accessed March 14th, 2013].
137. Campo, E., Miquel. R., Jares, P. *et al*. (1994). Prognostic significance of the loss of heterozygosity of *Nm23H-H1* and *p53* genes in human colorectal carcinomas. *Cancer* 73, 2913–2921.
138. Martinez, J.A., Prevot, S., Mordlinger, B. *et al*. (1995). Overexpression of nm23-H1 and nm23-H2 genes in colorectal carcinomas and loss of nm23-H1 expression in advanced tumour stages. *Gut* 37, 712-720.
139. Lamb, R.F., Going, J.J., Pickford, I. *et al*. (1996). Allelic imbalance at *NME1* in microdissected primary and metastatic human colorectal carcinomas is frequent but not associated with metastasis to lymph nodes or liver. *Cancer Res* 56, 916–920.
140. Goldman, E. (1907). The growth of malignant disease in man and the lower animals with special reference to the vascular system. *Lancet* 2, 1236–1240.
141. Gullino, P. M. (1978). Angiogenesis and oncogenesis. *J Natl Cancer Inst* 61, 639–643.
142. Bouck, N., Stellmach, V. Hsu, S. C. (1996). How tumors become angiogenic. *Adv Cancer Res* 69, 135–174.
143. Kerbel, R.S. (2000). Tumor angiogenesis: past, present and the near future. *Carcinogenesis* 21, 505-515.

144. Carmeliet, P., Ferreira, V., Breier, G. *et al.* (1996). Abnormal blood vessel development and lethality in embryos lacking a single vascular endothelial growth factor allele. *Nature* 380, 435–439.
145. Ferrara, N., Alitalo, K. (1999). Clinical applications of angiogenic growth factors and their inhibitors. *Nat Med* 5, 1359–1364.
146. Tammela, T., Enholm, B., Alitalo, K. *et al*. (2005). The biology of vascular endothelial growth factors. *Cardiovasc Res* 65, 550-563.
147. Lohela, M., Bry, M., Tammela, T. *et al*. (2009). VEGFs and receptors involved in angiogenesis versus lymphangiogenesis. *Current Opinion in Cell Biology* 21, 154-165.
148. Lohela, M., Saaristo, A., Veikkola, T. *et al*. (2003). Lymph-angiogenic growth factors, receptors and therapies. *Thromb Haemost* 90, 167-184.
149. Karpanen, T., Alitalo, K. (2001). Lymphatic vessels as targets of tumor therapy? *J Exp Med* 194, 37-42.
150. Pepper, M.S. (2001). Lymph-angiogenesis and tumor metastasis: myth or reality? *Clin Cancer Res* 7, 462-468.
151. Stacker, S.A., Caesar, C., Baldwin, M.E. *et al*. (2001). VEGF-D promotes the metastatic spread of tumor cells via the lymphatics. *Nat Med* 7, 186–191.
152. Von Marschall, Z., Scholz, A., Stacker, S.A. *et al.* (2005). Vascular endothelial growth factor-D induces lymphangiogenesis and lymphatic metastasis in models of ductal pancreatic cancer. *Int J Oncol* 27, 669–679.
153. Ferrara, N., Gerber, H.P., LeCouter, J. (2003). The biology of VEGF and its receptor. *Nat Med* 9, 669-676.
154. Achen, M.G., Stacker, S.A. (2006). Tumor lymph-angiogenesis and metastatic spread-New player begin to emerge. *Int J Cancer* 119, 1755-1760.
155. Toi, M., Hoshina, S., Takayanagi, T. *et al*. (1994). Association of vascular endothelial growth factor expression with tumor angiogenesis and with early relapse in primary breast cancer. *Jpn J Cancer Res* 85, 1045-1049.

156. Olofsson, B., Pajusola, K., Kaipainen, A. *et al*. (1996). Vascular endothelial growth factor B, a novel growth factor for endothelial cells. *Proc Natl Acad Sci USA* 93, 2576-2581.

157. Neufeld, G., Kessler, O., Herzong, Y. (2002). The interaction of Neuropilin-1 and Neuropilin-2 with tyrosine-kinase receptors for VEGF. *Adv Exp Med Biol* 515, 81-90.

158. Yla-Herttuala S, Rissanen TT, Vajanto I. *et al*. (2007). Vascular endothelial growth factors*:* biology and current status of clinical applications in cardiovascular medicine. *J Am Coll Cardiol* 49, 1015-1026.

159. Andre, T., Kotelevets, L. Vaillant, J.C. *et al*. (2000). Vegf, Vegf-B, Vegf-C and their receptors KDR, FLT-1 and FLT-4 during the neoplastic progression of human colonic mucosa. *Int J Cancer* 86, 174-181.

160. Niki, T., Iba, S., Tokunou, M. *et al*. (2000). Expression of vascular endothelial growth factors A, B, C, and D and their relationships to lymph node status in lung adenocarcinoma. *Clin Cancer Res* 6, 2431-2439.

161. Donnini, S., Machein, M.R., Plate, K. H. *et al*. (1999). Expression and localization of placenta growth factor and PIGF receptors in human meningiomas. *J Pathol* 189, 66-71.

162. Joukov, V., Pajusola, K., Kaipainen, A. *et al*. (1996). A novel vascular endothelial growth factor, VEGF-C, is a ligand for the FIt4 (VEGFR-3) and KDR (VEGFR-2) receptor tyrosine kinases. *EMBO J* 15, 290-298.

163. Chen, J.C., Chang, Y.W., Hong, C.C. *et al*. (2013). The role of the VEGF-C/VEGFRs axis in tumor progression and therapy. *Int J Mol Sci* 14, 88-107.

164. Mandriota, S.J., Jussila, L., Jeltsch, M. *et al*. (2001). Vascular endothelial growth factor-c-mediated lymph-angiogenesis promotes tumour metastasis. *EMBO J* 20, 672-682.

165. Arinaga, M., Noguchi, T., Takeno, S. *et al*. (2003). Clinical significance of vascular endothelial growth factor C and vascular endothelial growth factor receptor 3 in patients with non-small cell lung carcinoma. *Cancer* 97, 457-464.

166. Gu, Y., Qi, X., Guo, S. (2008). Lymph-angiogenesis induced by VEGF-C and VEGF-D promotes metastasis and a poor outcome in breast carcinoma: a retrospective study of 61 cases. *Clin Exp Metastasis* 25, 717-725.
167. Miyazaki, T., Okada, N., Ishibashi, K. *et al*. (2008). Clinical significance of plasma level of vascular endothelial growth factor-C in patients with colorectal cancer. *Jpn J Clin Oncol* 38, 839-843.
168. Achen, M.G., Jeltsch, M., Kukk, E. *et al*. (1998). Vascular endothelial growth factor D (VEGF-D) is a ligand for the tyrosine kinases VEGF receptor 2 (Flk1) and VEGF receptor 3 (Flt4). *Proc Natl Acad Sci USA* 95, 548-553.
169. Akahane, M., Akahane, T., Matheny, S.L. *et al.* (2006). Vascular endothelial growth factor-D is a survival factor for human breast carcinoma cells. *Int J Cancer* 118, 841-849.
170. Kleespies, A., Bruns, C.J., Jauch, K.W. (2005). Clinical significance of VEGF-A, -C and –D expression in esophageal malignancies. *Onkologie* 28, 281-288.
171. Ishii, H., Yazawa, T., Sato, H. *et al*. (2004). Enhancement of pleural dissemination and lymph node metastasis of intra-thoracic lung cancer cells by vascular endothelial growth factors (VEGFs). *Lung Cancer* 45, 325-337.
172. White, J.D., Hewett, P.W., Kosuge, D. *et al*. (2002). Vascular endothelial growth factor-D expression is an independent prognostic marker for survival in colorectal carcinoma. *Cancer Res* 62, 1669–1675.
173. Yokoyama, Y., Charnok-Jones, D.S., Licence, D. *et al.* (2003). Vascular endothelial growth factor-D expression is an independent prognostic marker for survival in epithelial ovarian carcinoma. *Br J Cancer* 88, 237-244.
174. Schimanski, C., Schlaegel, F., Jordan, M. *et al*. (2011). VEGF-D correlates with metastatic disease in gastric cancer patients undergoing surgery. *World J Surgery* 35, 1010-1016.
175. Bo, C., Xiaopeng, D., Chuanliang, P. *et al.* (2009). Expression of vascular endothelial growth factors C and D correlates with lymphangiogenesis and lymph node metastasis in lung adenocarcinoma. *Thorac Cardiovasc Surg* 57, 291-294.

176. Maglione, D., Guerriero, V., Viglietto, G. *et al*. (1991). Isolation of a human placenta cDNA coding for a protein related to the vascular permeability factor. *Proc Natl Acad Sci USA* 88, 9267-9271.

177. Persico, M.G., Vincenti, V., DiPalma, T. (1999). Structure, expression and receptor-binding properties of placenta growth factor (PIGF). L. Claesson-Welsh (Ed.), vascular growth factors and angiogenesis, *Springer-Verlag, Berlin* 31-40.

178. Nagy, J.A., Dvorak, A.M., Dvorak, H.F. (2003). VEGF-A (164/165) and PIGF: roles in angiogenesis and arteriogenesis. *Trends Cardiovasc Med* 13, 169-175.

179. Autiero, M., Waltenberger, J., Communi, D. *et al*. (2003). Role of PIGF in the intra- and intermolecular cross talk between the VEGF receptors FIt1 and FIk1. *Nat Med* 9, 936-943.

180. Fischer, M., Mazzone, B., Jonckx, P. *et al*. (2008). FLT1 and its ligands VEGFB and PIGF: drug targets for antiangiogenic therapy? *Nat Rev Cancer* 8, 942-956.

181. Wei, S.C., Liang, J.T., Tsao, P.N. *et al*. (2009). Preoperative serum placenta growth factor level is a prognostic biomarker in colorectal cancer. *Dis Colon Rectum* 52, 1630-1636.

182. Chen, C.N., Hsieh, F.J., Cheng, Y.M. *et al*. (2004). The significance of placenta growth factor factor in angiogenesis and clinical outcome of human gastric cancer. *Cancer Lett* 213, 73-82.

183. Parr, C., Watkins, G., Boulton, M. *et al*. (2005). Placenta growth factor is over-expressed and has prognostic value in human breast cancer. *Eur J Cancer* 41, 2819-2827.

184. Shibuya, M., Yamaguchi, S., Yamane, A. *et al*. (1990). Nucleotide sequence and expression of a novel human receptor-type tyrosine kinase gene (flt), closely related to the fms family. *Oncegene* 5, 519-524.

185. Kato, H., Yoshikawa, M., Miyuzaki, T. *et al.* (2002). Expression of vascular endothelial growth factor (VEGF) and its receptors (FIt-1 and FIk-1) in esophageal squamous cell carcinoma. *Anticancer Res* 22, 3977-3984.

186. Jackson, M.W., Roberts, S., Heckford, S.E. *et al.* (2002). A potential autocrine role for vascular endothelial growth factor in prostate cancer. *Cancer Res* 62, 854-859.

187. Casalou, C., Costa, A., Carvalho, T. *et al*. (2011). Cholestrol regulate VEGFR-1(*FLT-1*) expression and signaling in acute leukemia cells. *Mol Cancer Res* 9, 215-224.

188. Wei, S.C, Tsao, P.N., Weng, M.T. *et al*. (2013). Flt-1 in colorectal cancer cells is required for the tumor invasive effects of placental growth factor through a p38-MMP9 pathway. *J Biomed Sci* 20-2-12.

189. Terman, B.I., Carrion, M.E., Kovacs, E. *et al*. (1991). Identification of a new endothelial cell growth factor receptor tyrosine kinase. *Oncogene* 6, 1677-1683.

190. Shibuya, M. (2006). Differential roles of vascular endothelial growth factor receptor-1 and receptor-2 in angiogenesis. *J Biochem Mol Biol* 39, 469-478.

191. Fine, B.A., Valente, P.T., Feinstein, G.I. *et al*. (2000). VEGF, flt-1, and KDRÉflk-1 as prognostic indictors in endometrial carcinoma. *Gynecol* 76, 33-39.

192. Martins, S.F., Garcia, E.A., Luz, M.A. *et al*. (2013). Clinicopathological correlation and prognostic significance of VEGF-A VEGFR-1 and VEGR-3 expression in colorectal cancer. *Cancer Genomics Proteomics* 10, 55-67.

193. Jennbacken, K., Vallbo, C., Wang, W. *et al*. (2005). Expression of vascular endothelial growth factor C (VEGF-C) and VEGF receptor-3 in human prostate cancer is associated with regional lymph node metastasis. *Prostate* 65, 110–116.

194. Valtola, R., Salven, P., Heikkila, P. *et al*. (1999). VEGFR-3 and its ligand VEGF-C are associated with angiogenesis in breast cancer. *Am J Pathol* 154, 1381–1390.

195. Achen, M.G., Williams, R.A., Minekus, M.P. *et al*. (2001). Localization of vascular endothelial growth factor-D in malignant melanoma suggests a role in tumour angiogenesis. *J Pathol* 193, 147–154.

196. Weidner, N., Semple, J.P., Welch, W.R. *et al*. (1991). Tumor angiogenesis and metastasis .correlation in invasive breast carcinoma. *N Engl J Med 324*, 1-8.

197. Weidner, N., Carroll, P.R., Flax, J. *et al*. (1993). Tumor angiogenesis correlate with metastasis in invasive prostate carcinoma. *Am J Pathol* 143, 401-409.

198. Yamakazi, K., Abe, S., Tekekawa, H. *et al*. (1994). Tumor angiogenesis in human lung adenocarcinoma. *Cancer* 74, 2245-2250.

199. Maeda, K., Chung, Y.S., Takasuka, S. *et al*. (1995). Tumor angiogenesis as a predictor of recurrence in gastric carcinoma. *J Clin Oncol* 13, 477-481.

200. Wiggnins, D.L., Granai, C.O., Steinhoff, M.M. *et al*. (1995). Tumor angiogenesis as a prognostic factor in cervical carcinoma. *Gynecol Oncol* 56, 353-356.

201. Lu, P., Weaver, V.M., Werb, Z. (2012). The extracellular matrix: a dynamic niche in cancer progression. *J Cell Biol 196*, 395-405.

202. Aznavoorian, S., Murphy, A.N., Stetler-Stevenson, W.G. *et al*. (1993). Molecular aspects of tumor cell invasion and metastasis. *Cancer* 71, 1368-1383.

203. Stetler-Stevenson, W.G., Aznavoorian, S., Liotta, L.A. (1993). Tumor cell interactions with the extracellular matrix during invasion and metastasis. *Ann Rev Cell Biol* 9, 541-573.

204. Chambers, A.F., Matrisian, L.M. (1997). Changing views of the role of matrix metalloproteinases in metastasis. *J NatI Cancer Inst* 89, 1260-1270.

205. Gray, K.A., Daugherty, L.C., Gordon, S.M. *et al*. (2013). Genenames.org: the HGNC resources in 2013. *Nucleic Acids Res* 41, 545-552.

206. Nikkola, J., Vihinen, P., Vuoristo, M.S. *et al*. (2005). High serum levels of matrix metalloproteinase-9 and matrix metalloproteinase-1 are associated with rapid progression in patients with metastatic melanoma. *Clin Cancer Res* 11, 5158-5166.

207. Xie, M., Sun, Y., Li, Y. (2004). Expression of matrix metalloproteinases in supraglottic carcinoma and its clinical implication for estimating lymph node metastases. *Laryngoscope* 114, 2243-2248.

208. Katayama, A., Bandoh, N., Kishibe, K. *et al*. (2004). Expressions of matrix metalloproteinases in early stage oral squamous cell carcinoma as predictive indictors for tumor metases and prognosis. *Clin Cancer Res* 10, 634-640.

209. Ishigami, S.I., Arii, S., Furutani, M. *et al*. (1998). Predictive value of vascular endothelial growth factor (VEGF) in metastasis and prognosis of human colorectal cancer. *Br J Cancer* 78, 1379-1384.

210. Tokunaga, T., Oshika, Y., Abe, Y. *et al*. (1998). Vascular endothelial growth factor (VEGF) mRNA isoform expression pattern is correlated with liver metastasis and poor prognosis in colon cancer. *Br J Cancer* 77, 998–1002.

211. Takahashi, Y., Kitadai, Y., Bucana, C.D. *et al*. (1995). Expression of vascular endothelial growth factor and its receptor, KDR, correlates with vascularity, metastasis, and proliferation of human colon cancer. *Cancer Res* 55, 3964–3968.

212. Cascinu, S., Staccioli, M.P., Gasparini, G. *et al*. (2000). Expression of vascular endothelial growth factor can predict event-free survival in stage II colon cancer. *Clin Cancer Res* 6, 2803-2807.

213. Zafirellis, K., Agrogiannis, G., Zachaki, A. *et al*. (2008). Prognostic significance of VEGF expression evaluated by quantitative immunohistochemical analysis in colorectal cancer. *J Surg Res* 147, 99–107.

214. Werther, K., Christensen, I.J., Brunner, N. *et al*. (2000). Soluble vascular endothelial growth factor levels in patients with primary colorectal carcinoma. The Danish RANX05 Colorectal Cancer Study Group. *Eur J Surg Oncol* 267, 657–662.

215. De Vita, F., Orditura, M., Lieto, E. *et al*. (2004). Elevated perioperative serum vascular endothelial growth factor levels in patients with colon carcinoma. *Cancer* 100, 270-278.

216. Wei, S.C., Tsao, P.N., Yu, S.C. *et a*l. (2005). Placenta growth factor expression is correlated with survival of patients with colorectal cancer. *Gut* 54, 666-672.

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

218. White, J.D., Hewett, P.W., Kosuge, D. *et al*. (2002). Vascular endothelial growth factor-D expression is an independent prognostic marker for survival in colorectal carcinoma. *Cancer Res* 62, 1669–1675.

219. Rmali, K.A., Puntis, M.C.A., Jiang, W.G. (2006). Level of the expression of VEGF-A, B, C, D, and their receptors (FLT-1, KDR, and FLT-4) and its correlation with prognosis in patients with colorectal cancer. *Int J Cancer Res* 2, 31-41.

220. Vacirca, Z.S. (2004). Role of matrix metalloproteinases (MMPs) in colorectal cancer. *Cancer Metastasis Rev* 23, 101-117.

221. Hinoda, Y., Okayama, N., Takano, N. *et al*. (2002). Association of functional polymorphisms of matrix metalloproteinase (MMP)-1 and MMP-3 genes with colorectal cancer. *Int J Cancer* 102, 526-529.

222. Li, C.Y., Yuan, P., Lin, S.S. *et al*. (2013). Matrix metalloproteinase 9 expression and prognosis in colorectal cancer: a meta-analysis. *Tumour Biol* 34, 735-741.

223. Langers, A.M.J., Sier, C.F.M, Hawinkles L.J.A.C. *et al*. (2008). MMP-2 geno-phenotype is prognostic for colorectal cancer survival, whereas MMP-9 is not. *Br J Cancer* 98, 1820–1823.

224. Shi, M., Yu, B., Gao, H. *et al*. (2013). Matrix metalloproteinase 2 over-expression and prognosis in colorectal cancer: a meta-analysis. *Mol Biol Rep* 40, 617-623.

225. Yang, W., Arii, S., Gorrin-Rivas, M.J. *et al*. (2001). Human macrophage metalloelastase gene expression in colorectal carcinoma and its clinicopathologic significance. *Cancer* 91, 1277-1283.

226. Savas, S., Younghusband B. (2010). dbCPCO: A database of genetic markers tested for their predictive and prognostic value in colorectal cancer. *Hum Mutat* 31, 901-907.

227. Dassoulas, K., Gazouli, M., Rizos, S. *et al.* (2008). Common polymorphisms in the vascular endothelial growth factor gene and colorectal cancer development, prognosis, and survival. *Mol Carcinog* 48, 563-569.

228. Lurje, G., Zhang, W., Schulthesis, A.M. *et al*. (2008). Polymorphisms in VEGF and IL-8 predict tumor recurrence in stage III colon cancer. *Ann Oncol* 19, 1734-1741.

229. Yamamori, M., Taniguchi, M., Maeda, S. *et al*. (2008). VEGF T-1498C polymorphism, a predictive marker of differentiation of colorectal adenocarcinomas in Japanese. *Int J Med Sci* 5, 80-86.

230. Lurje, G., Hendifar, A.E., Schulthesis, A.M. *et al.* (2009). Polymorphisms in interleukin 1 beta and interleukin 1 receptor antagonist associated with tumor recurrence in stage II colon cancer. *Pharmacogenet Genomics* 19, 95-102.

231. Loupakis, F., Ruzzo, A., Salvatore, L. *et al*. (2011). Retrospective exploratory analysis of VEGF polymorphisms in the prediction of benefit from first-line FOLFIRI plus bevacizumab in metastatic colorectal cancer. *BMC Cancer* 11, 2-9.
232. Gerger, A., El-Khoueiry, A., Zhang, W. *et al*. (2011). Pharmacogenetic angiogenesis profiling for first-line Bevacizumab plus oxaliplatin-based chemotherapy in patients with metastatic colorectal cancer. *Clin Cancer Res* 17, 5783-5792.
233. Zhang, W., Gordon, M., Press, O.A. *et al.* (2006). Cyclin D1 and epidermal growth factor polymorphisms associated with survival in patients with advanced colorectal cancer treated with Cetuximab. *Pharmacogenet Genomics* 16, 475-483.
234. Chen, M.H., Tzeng, C.H., Chen, P.M. *et al*. (2011). VEGF -460T> C polymorphism and its association with VEGF expression and outcome to FOLFOX-4 treatment in patients with colorectal carcinoma. *Pharmacogenomics J* 11, 227-236.
235. Hansen, T.F., Garm Spindler, K.L., Andersen, R.F. *et al*. (2011). The predictive value of genetic variations in the vascular endothelial growth factor A gene in metastatic colorectal cancer. *Pharmacogenomics J* 11, 53-60.
236. Hasen, T.F., Sorense, F.B., Spindler, K.L. *et al*. (2010). Microvessel density and the association with single nucleotide polymorphism of the vascular endothelial growth factor receptor 2 in patients with colorectal cancer. *Vircgows Arch* 456, 251-260.
237. Pander, J., Wessels, J.A., Gelderblom, H. *et al*. (2011). Pharmacogenetic interaction analysis for the efficacy of systemic treatment in mestastic colorectal cancer. *Ann Oncol* 22, 1147-1153.
238. Hansen, T.F., Christensen, R.D., Andersen, R.F. *et al*. (2012). The predictive value of single nucleotide polymorphisms in the VEGF system to the efficacy of first-line treatment with bevacizumab plus chemotherapy in patients with metastatic colorectal cancer cancer: results from the Nordic ACT trial. *Int J Colorectal Dis* 27, 715-720.

239. Hettiaratchi, A., Hawkins, N.J., McKenzie, G. *et al*. (2007). The collagenase-1 (MMP-1) gene promoter polymorphism - 1607/2G is associated with favourable prognosis in patients with colorectal cancer. *Br J Cancer* 96, 783-792.

240. Ghilardi, G., Biondi, M.L., Mangoni, J. *et al.* (2001). Matrix metalloproteinase-1 promoter polymorphism 1G/2G is correlated with colorectal cancer invasiveness. *Clin Cancer Res* 7, 2344-2346.

241. Woo, M., Park, K., Nam, J. *et al*. (2007). Clinical implications of matrix metalloproteinase-1, -3, -7, -9, -12, and plasminogen activator inhibitor-1 gene polymorphisms in colorectal cancer. *J Gastroenterol Hepatol* 22, 1064-1070.

242. Elander, N., Soderkvist, P., Fransen, K. (2006). Matrix metalloproteinase (MMP) -1, -2, -3 and -9 promoter polymorphisms in colorectal cancer. *Anticancer Res* 26, 791-795.

243. Langers, A.M., Sier, C.F., Hawinkels, L.J. *et al*. (2008). MMP-2 geno-phenotype is prognostic for colorectal cancer survival, whereas MMP-9 is not. *Br J Cancer* 98, 1820-1823.

244. Xu, E., Lai, M., Lv, B. *et al.* (2004). A single nucleotide polymorphism in the matrix metalloproteinase-2 promoter is associated with colorectal cancer. *Biochem Biophys Res Commun* 324, 999-1003.

245. Kang, M.J., Jung, S.A., Jung, J.M. *et al*. (2011). Associations between single nucleotide polymorphisms of MMP2, VEGF, and HIF1A genes and the risk of developing colorectal cancer. *Anticancer Res* 31, 575-584.

246. Cargill, M., Altshuler, D., Ireland, J. *et al*. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 22, 231-238.

247. Scott, L.J., Athey, B., Watson, S.J. *et al*. (2006). SNP functional portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics* 22, 523-529.

248. Rebbeck, R., Ambrosone, C.B., Bell, D.A. *et al*. (2004). SNPs, haplotypes, and cancer: applications in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev* 13, 681-687.

249. Melzer, D. (2008). Genetic polymorphisms and human aging: association studies deliver. *Rejuvenation Res* 11, 523–526.

250. Kim, D.S. (2007). Thyroid cancer: are molecular studies making any difference? *J Laryngol Otol* 121, 917-926.

251. Johnson, A.D., O'Donnell, C.J. (2009). An open access database of genome-wide association results. *BMC Med Genet* 10, 1-17.

252. Zhao, H., Pfeiffer, R., Gail, M.H. (2003). Haplotype analysis in population genetics and association studies. *Pharmacogenomics* 4, 171-178.

253. Ardlie, K., Kruglyak, L., Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 3, 299–309.

254. Weiss, K. M., Clark, A. G. (2002). Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 18, 19–24.

255. Zhang, K., Calabrese, P., Nordborg, M. *et al*. (2002). Haplotype block structure and its application to association studies: power and study designs. *Am J Hum Genet* 71, 1386-1394.

256. Daly, M.J., Rioux, J.D., Schaffner, S.F. *et al*. (2001). High resolution haplotype structure in the human genome. *Nat Genet* 29, 229-232.

257. Jin, A.L., Xiong, M. (2001). Haplotypes vs single marker linkage disequilibrium tests: what do we gain? *Eur J Hum Genet* 9, 291-300.

258. Crawford, D.C., Nickerson, D.A. (2005). Definition and clinical importance of haplotypes. *Annu Rev Med* 56, 303-320.

259. Excoffier, L., Slatkin, M. (1995). Maximum likelihood estimation of molecular haplotype frequencies is a diploid population. *Mol Biol Evol* 12, 921–927.

260. Excoffier, L., Laval, G., Balding, D. (2003). Gametic phase estimation over large genomic regions using an adaptive window approach. *Hum Genomics* 1, 7–19.

261. Stephens, M., Smith, N.J., Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68, 978–989.

262. Stephens, M., Donnelly, P. (2003). A comparison of Bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73, 1162–1169.

263. Niu, T., Qin, Z.S., Xu, X. *et al*. (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 70, 157–169.

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

265. Negandhi, A.A., Hyde, A., Dick, E. *et al*. (2013). MTHFR Glu429Ala and ERCC5 His46His polymorphisms are associated with prognosis in colorectal cancer patients: analysis of two independent cohorts from Newfoundland. *Plos One* 8(4): e61469.

266. Wood, M.O., Younghusband, H.B., Parfrey, P.S. *et al*. (2010). The genetic basis of colorectal cancer in a population-based incident cohort with a high rate of familial disease. *Gut* 59, 1369-1377.

267. The genomic coverage of Illumina® human Omni1-Quad genome-wide platform http://www.dkfz.de/gpcf/fileadmin/downloads/Genotyping/datasheet_infiniumhd.pdf [Access May 21, 2013].

268. Turner, S., Armstrong, L.L., Bradford, Y. *et al*. (2011). Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet* 68, 1-19.

269. Carroll, J.D., Arabie, P. (1980). Multidimensional scaling. Rosenzweig, M.R. and Porter, L.W., eds. *Annu Review Psychol* 31, 607-649.

270. Ringner, M. (2008). What is principal component analysis? *Nature Biotech* 26, 303-304.

271. Jolliffe, I.T. Principal Component Analysis (Springer, New York, 2002).

272. Kent, W.J., Sugnet, C.W., Furey, T.S. *et al*. (2002). The human genome browser at UCSC. *Genome Res* 12, 996-1006.

273. Meyer, L.R., Zweig, A.S., Hinrichs, A.S. *et al*. (2012). The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res* 41, 64-69.

274. Purcell, S., Neale, B., Todd-Brown, K. *et al*. (2007). PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Hum Genet* 81, 559-575.

275. Purcell, S. http://pngu.mgh.harvard.edu/purcell/plink/ (Accessed January 2013).

276. Sherry, S.T., Ward, M., Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* 9, 677-679.

277. Savas, S., Liu, G., Xu, W. (2013). Special considerations in prognostic research in cancer involving genetic polymorphisms. *BMC Med* 11: 149.

278. Dai, J., Gu, J., Huang, M. *et al*. (2012). GWAS-identified colorectal cancer susceptibility loci associated with clinical outcome. *Carcinogenesis* 33, 1327-1331.

279. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc 1958; 53:457-81.

280. Stephens M. http://stephenslab.uchicago.edu/software.html#phase [Accessed May 2013].

281. Cox, D. (1972). Regression models and life tables (with discussion). *J Roy Stat Soc Series B* 34, 187-220.

282. Field, A. (2009). Discovering statistics using SPSS (3rd Edition). SAGE, London, Uk.

283. http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/haploview/haploview [Accessed October 2013].

284. Hicklin, D.J Ellis, L.M. (2005). Role of the vascular endothelial growth factor pathway in tumor growth and angiogenesis. *J Clin Oncol* 23, 1011-1027.

285. Bergers, G., Brekken, R., McMachon, G. *et al*. (2000). Matrix metalloproteinase-9 triggers the angiogeneic switch during carcinogenesis. *Nat Cell Biol* 2, 737-744.

286. Renner, W., Kotschan, S., Hoffman, C. *et al*. (2000). A common 936C/T mutation in the gene for vascular endothelial growth factor is associated with vascular endothelial factor plasma level. *J Vasc Res* 37, 443-448.

287. Stevens, A., Sodden, J., Brenchley, P.E. *et al*. (2003). Haplotype analysis of the polymorphic human vascular endothelial growth factor gene promoter. *Cancer Res* 63, 812-816.

288. Fang, S., Jin, X., Wang, R. *et al*. (2005). Polymorphisms in the MMP1 and MMP3 promoter and non-small cell lung carcinoma in North China. *Carcinogenesis* 26, 481-486.

289. Matsumura, S., Oue, N., Nakayama, H. *et al*. (2005). A single nucleotide polymorphism in the MMP-9 promoter affects tumor progression and invasive phenotype of gastric cancer. *J Cancer Res Clin Oncol* 131, 19-25.

290. Six, L., Grimm, C., Leodolter, S. *et al*. (2006). A polymorphism in the matrix metalloproteinase-1 gene promoter is associated with the prognosis of patients with ovarian cancer. *Gynecol Oncol* 100, 506-510.

291. Emmanouilides, C., Pegram, M., Robinson, R. *et al*. (2004). Anti-VEGF antibody bevacizumab (Avastin) with 5FU/LV as third line treatment for colorectal cancer. *Tech Coloproctol* 8, 50-52.

292. Yakes, F.M., Chen, J., Tan, J. *et al*. (2011). Cabozantinib (XL184), a novel MET and VEGFR2 inhibitor, simultaneously suppresses metastasis, angiogenesis, and tumor growth. *Mol Cancer Ther* 10, 2298-2308.

293. Xu, Z., Taylor, J.A. (2008). SNPinfor: integrating GWAS candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res* 37, 600-605.

294. Entrez Gene: MMP8 matrix metallopeptidase 8 (neutrophil collagenase) http://www.ncbi.nlm.nih.gov/gene/4317 [Accessed March 17th, 2014].

295. Entrez Gene: MMP27 matrix metallopeptidase 27 [homo sapiens (human)] http://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=64066, http://www.uniprot.org/uniprot/Q9H306#section_comments [Accessed March 17th, 2014].

296. Murugan, A.K., Dong, J., Xie, J. *et al*. (2011). Uncommon GNAQ, MMP8, AKT3, EGFR, and PIK3R1 mutations in thyroid cancers. *Endocr Pathol* 22, 97-102.

297. Palavalli, L.H., Prickett, T.D., Wunderlich, J.R. *et al*. (2009). Analysis of the matrix metalloproteinase family reveals that MMP8 is often mutated in melanoma. *Nat Genet* 41, 518-520.

298. Neff, R.L., Farrar, W.B., Kloos, R.T. *et al*. (2008). Anaplastic thyroid cancer. *Endocrino Metab Clin North Am* 37, 525-538.

299. Entrez Gene: MMP3 matrix metallopeptidase 3 (stromelysin 1, progelatinase) [homo sapiena (human)] http://www.ncbi.nlm.nih.gov/gene/4314.[Accessed March 24th, 2014].
300. Hinoda, Y., Okayama, N., Takano, N. *et al*. (2002). Association of functional polymorphisms of matrix metalloproteinase (MMP)-1 and MMP-3 genes with colorectal cancer. *Int J Cancer* 102, 526-529.
301. Wang, J., Xu, D., Wu, X. *et al*. (2011). Polymorphisms of matrix metalloproteinases in myocardial infarction: a meta-analysis. *Heart* 97, 1542-1546.
302. Matsuyama, A., Sakai, N., Ishigami, M. *et al*. (2003). Matrix metalloproteinases as novel diease markers in takayasu arteritis. *Circulation* 108, 1469-1473.
303. Flex, A., Giovannini, S., Biscetti, F. *et al*. (2013). Effect of proinflammatory gene polymorphisms on the risk of alzheimer's disease. *Neurodegener Dis* 13, 230-236.
304. Dey, S., Stalin, S., Gupta, A. *et al*. (2012). Matrix metalloproteinase3 gene promoter polymorphisms and their haplotypes are associated with gastric cancer risk in eastern Indian population. *Mol Carcinog* 51, 42-53.
305. Pendas, A.M., Santamaria, I., Alvarez, M.V. *et al*. (1996). Fine physical mapping of the human matrix metalloproteinase genes clustered on chromosome 11q22.3. *Genomics* 37, 266-268.
306. Haja Mohideen, A.M.S., Dicks, E., Parfrey, P. *et al*. Mitochondrial DNA polymorphisms, its copy number change and outcome in colorectal cancer (submitted to BMC Research Notes).
307. Haja Mohideen, A.M.S., Hyde, A., Squires, J. *et al*. Examining the polymorphisms in the hypoxia pathway genes in relation to survival outcomes in colorectal cancer (submitted to PLOS ONE).

Appendices

Appendix A: 381 polymorphisms investigated in this study

Genes	Polymorphisms			
VEGFA	rs2010963	rs3024994		
	rs25648	rs2146323		
	rs833068	rs3025010		
	rs833069	rs3025035		
	rs833070	rs3025039		
	rs3025053			
VEGFB	rs11603042			
	rs4930152			
VEGFC	rs2877961	rs2171083	rs3775202	rs10012721
	rs17697359	rs1564922	rs11947611	rs13122901
	rs1485762	rs1485768	rs3775198	rs4557213
	rs7664413	rs6820170	rs3775195	rs10000057
	rs1485766	rs475106	rs2333526	
PGF	rs8185			
	rs12411			
VEGFR1	rs9554314	rs7332329	rs7324547	rs585421
	rs12429309	rs9508021	rs17086609	rs7323184
	rs9513070	rs2104330	rs1853581	rs622227
	rs12877323	rs9319427	rs7989623	rs675923
	rs3794397	rs9319429	rs7995976	rs655024
	rs3794399	rs9513099	rs1408243	rs679791
	rs2296188	rs10507384	rs9551462	rs600640
	rs2296189	rs9513105	rs9554325	rs598945
	rs7987291	rs11149523	rs3751395	rs17537350
	rs7987649	rs9508034	rs17086617	rs3794405
	rs942364	rs9513112	rs2387632	rs9513113
	rs3794400	rs9554330	rs3936415	rs10507386
	rs1324057			
VEGFR2	rs12642307	rs2034965	rs6828477	
	rs2125489	rs17711073	rs2168945	
	rs1531289	rs11941492	rs11732292	
	rs17709898	rs2305948	rs1870377	
	rs17085265	rs7692791	rs17085326	
	rs2219471	rs6837735		

	rs6838752	rs12502008	rs3797104	
VEGFR3	rs307822	kgp53910	rs307823	
	rs2279622	rs2290983	rs3797102	
	rs11739750	rs10085025	rs3736061	
	rs11747066	rs4700745		
	rs10058772	rs10072977		
	rs2242217	rs307806		
	rs400330	rs11748431		
	rs1130378	rs307814		
MMP1	rs5854	rs7125062		
	rs2071230	rs470558		
	rs2239008	rs10488		
	rs470215	rs3213460		
	rs470747			
	rs1938901			
MMP2	rs1477017	rs1992116	rs2287074	
	rs865094	rs2287076	rs243843	
	rs17301608	rs11639960	rs243842	
	rs1132896	rs243836	rs183112	
	rs1053605	rs243835		
	rs866770	rs243834		
	rs9302671	rs10775332/rs14070		
	rs2241145	rs11541998		
	rs243845	rs7201		
MMP3	rs566125			
	rs3025066			
	rs3020919			
	rs679620			
MMP7	rs17886371			
	rs14983			
	rs2156528			
	rs1996352			
	rs10502001			
MMP8	rs12365082			
	rs7934972			
	rs12284255			
	rs3740938			
	rs2012390			
	rs1940475			
	rs6590984			

	rs3765620			
	rs2155052			
MMP9	rs2274755			
	rs17576			
	rs2236416			
	rs2274756/rs17577			
	rs13925			
	rs20544			
MMP10	rs470168	rs4431992		
	rs17293348	rs2276108		
	rs470171	rs17860950		
	rs12290253	rs17293607		
	rs547561	rs486055		
	rs12272341			
MMP11	rs738791			
	rs2267029			
	rs738792			
MMP12	rs17368582			
	rs11225442			
	rs7123600			
MMP13	rs10895372			
	rs10502009			
	rs3819089			
	rs640198			
MMP14	rs1042703			
	rs762052			
	rs8006914			
	rs17243048			
	rs2236302			
	rs1042704			
	rs2236307			
	rs743257			
	rs17882342			
MMP15	rs41522747			
	rs11648508			
	rs3743563			
	rs1050779			
MMP16	rs2664369	rs10089111	rs1879201	rs2664352
	rs2664370	rs9297422	rs17666490	rs11782395
	rs10097366	rs1382105	rs16880099	rs1477916

	rs2616496	rs16878625	rs4961082	rs17664125
	rs17719609	rs1477917	rs7826477	rs13277637
	rs16877270	rs2664361	rs6994019	rs16878008
	rs1477908	rs16878818	rs16880416	rs2616487
	rs10103111	rs10099888	rs1467251	rs6469206
	rs2616493	rs7819728	rs10955542	rs7826929
	rs10098052	rs1996637	rs2222294	rs2616506
	rs2664346	rs1519938	rs7817382	rs17663841
	rs2616488	rs6981717	rs10100297	rs977231
	rs13261974	rs13256568	rs7835845	rs7000030
	rs6469298	rs2176771	rs9771895	rs3851539
	rs17666351	rs1519942	rs16878034	rs10504846
	rs13261169	rs12546847	rs4961076	rs10094702
	rs1401861	rs17722347	rs7834743	
	rs10504847	rs4961080	rs7816934	
MMP17	rs4964924	rs9634312		
	rs4964927	rs11613757		
	rs11246838	rs11835665		
	rs6598163	rs10751704		
	rs34515698	rs12099648		
	rs10751700	rs3087864		
	rs7300198			
MMP19	rs2242295			
	rs2291267			
	rs2291268			
MMP20	rs2292730	rs1784424	rs10895322	rs2280211
	rs11225332	rs1784423	rs1711430	rs11225344
	rs1711399	rs3781787	rs1784430	rs1962082
	rs1784439	rs3781788	rs1711427	rs2245803
	rs7116339	rs17098913	rs1784425	
	rs1711433	rs10502005		
MMP21	rs7922546			
	rs10901424			
	rs12775804			
MMP24	kgp4728036	kgp7633769	kgp420199	rs11696548
	kgp4471741	kgp9807173	rs6088776	kgp4501520
	kgp6966600	rs2425032	rs2247828	rs6060341
	kgp8495749	rs1205411	rs2425024	rs7280
	kgp481229	kgp1472099	kgp7289875	
	rs12479765	rs2254207	kgp5576338	

	rs2425022	kgp4265649	kgp10149373	
MMP25	rs2247226			
	rs10431961			
	rs7199221			
	rs1064875			
	rs1064948			
	rs11864930			
	rs10438593			
MMP26	rs2499958			
MMP27	rs2509010	rs17099425	rs2846703	
	rs11607205	rs11225386	rs3809018	
	rs1276289	rs11225388	rs4754870	
	rs11821641	rs2846707		
	rs1276286	rs1939015		
	rs2846723	rs12099177		
	rs2846701	rs11225389		
MMP28	rs3826404			
	rs10451309			

Appendix B: Results of the univariable analysis for clinicopathological and other features and overall survival

			95% CI for HR		
Variable	**p-value**	**HR**	**Lower**	**Upper**	**n**
Age at diagnosis	0.182	1.012	0.995	1.029	504
Sex (male vs female)	**0.017**	1.479	1.071	2.042	504
Histology (mucinous vs non-mucinous)	0.782	0.933	0.572	1.521	504
Location (rectum vs colon)	0.238	1.205	0.884	1.643	504
Stage	**<0.001**				504
Stage II vs stage I	0.222	1.42	0.809	2.493	
Stage III vs stage I	**0.003**	2.313	1.335	4.008	
Stage IV vs stage I	**<0.001**	9.925	5.544	17.766	
Grade (poorly differentiated vs well/moderately differentiated)	0.592	0.84	0.443	1.591	500
Vascular invasion (+ vs -)	**<0.001**	1.71	1.251	2.336	466
Lymphatic invasion (+ vs -)	**0.005**	1.564	1.148	2.132	464
Familial risk (high/intermediate risk vs low risk)	0.687	1.064	0.787	1.439	504
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.165	0.061	0.446	483
BRAF Val600Glu mutation (+ vs -)	0.258	0.72	0.407	1.273	457
Adjuvant chemotherapy status (+ vs -)	0.679	1.067	0.786	1.448	500
Adjuvant 5-FU based chemotherapy status (+ vs -)	0.975	1.005	0.738	1.368	490
Adjuvant radiotherapy status (+ vs -)	0.67	1.078	0.762	1.525	487

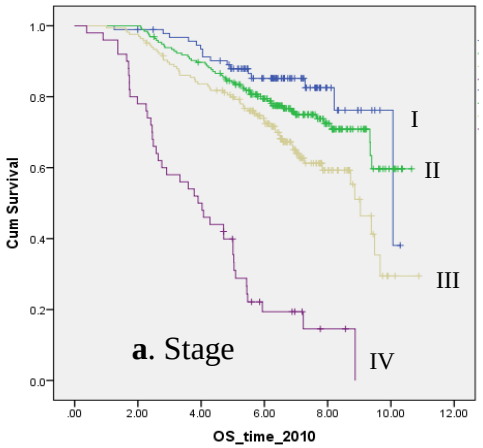
HR: hazard ratio, CI: confidence interval, n: number of patients, 5-FU 5-Fluorouracil, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable, (+) means present, (-) means absent.

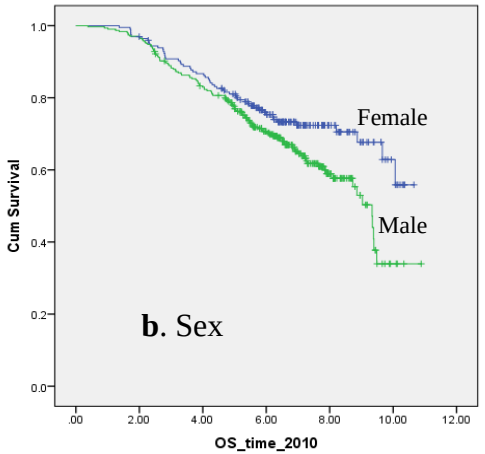
Appendix C: Results of the univariable analysis for clinicopathological and other features and disease-free survival

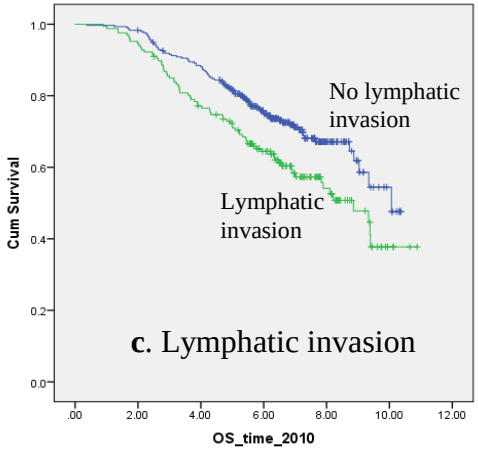
			95% CI for HR		
Variable	**p-value**	**HR**	**Lower**	**Upper**	**N**
Age at diagnosis	0.558	1.005	0.989	1.02	503
Sex(male vs female)	**0.014**	1.449	1.076	1.951	503
Histology(mucinous vs non-mucinous)	0.695	0.913	0.581	1.436	503
Location(rectum vs colon)	**0.035**	1.358	1.022	1.804	503
Stage	**<0.001**				503
Stage II vs stage I	0.221	1.363	0.83	2.24	
Stage III vs stage I	**<0.001**	2.345	1.445	3.804	
Stage IV vs stage I	**<0.001**	5.872	3.465	9.954	
Grade (poorly differentiated vs well/moderately differentiated)	0.418	0.778	0.423	1.429	499
Vascular invasion (+ vs -)	**<0.001**	1.651	1.236	2.205	465
Lymphatic invasion (+ vs -)	**0.003**	1.539	1.155	2.051	463
Familial risk (high/intermediate risk vs low risk)	0.296	1.16	0.878	1.533	503
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.254	0.119	0.54	482
BRAF Val600Glu mutation (+ vs -)	0.474	0.833	0.506	1.373	457
Adjuvant chemotherapy status (+ vs -)	0.299	1.162	0.876	1.542	499
Adjuvant 5-FU based chemotherapy status (+ vs -)	0.538	1.094	0.822	1.456	489
Adjuvant radiotherapy status (+ vs -)	0.146	1.262	0.922	1.727	486

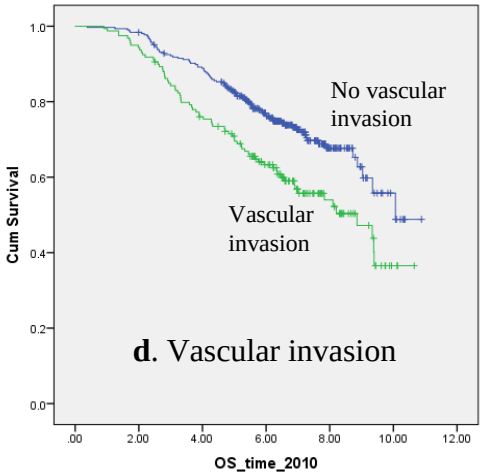
HR: hazard ratio, CI: confidence interval, n: number of patients, 5-FU 5-Fluorouracil, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable, (+) means present, (-) means absent.

Appendix D: Kaplan-Meier survival plots for the baseline variables associated with overall survival

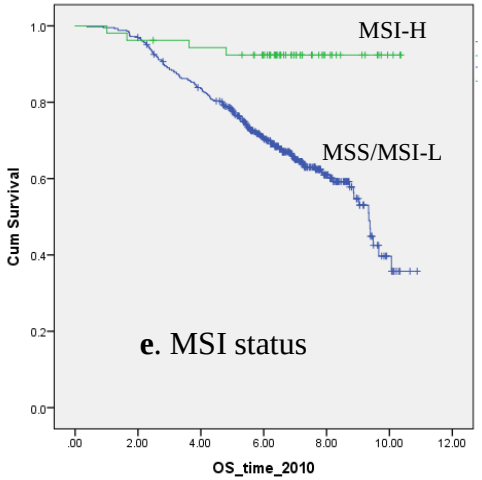






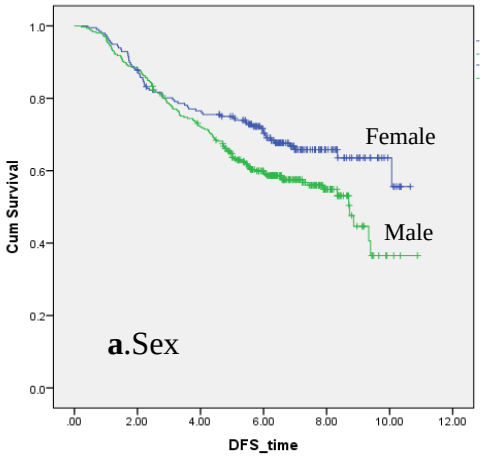


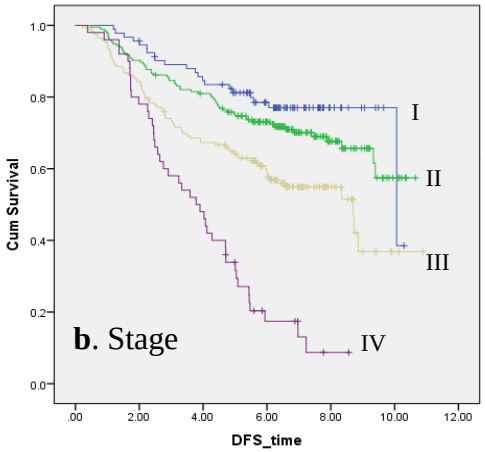
d. Vascular invasion

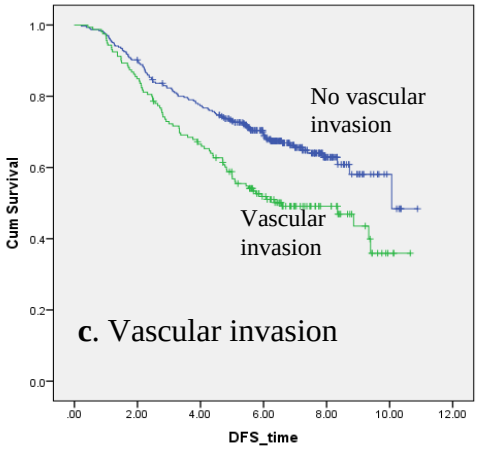


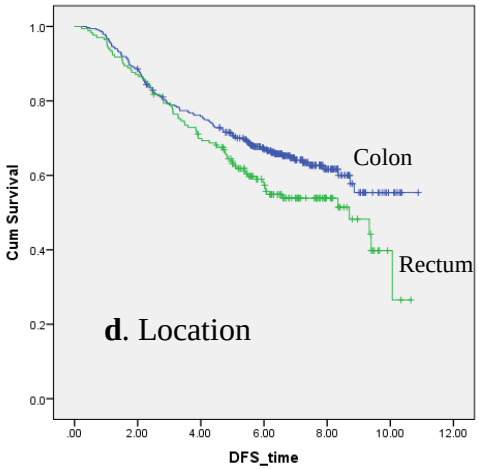
e. MSI status

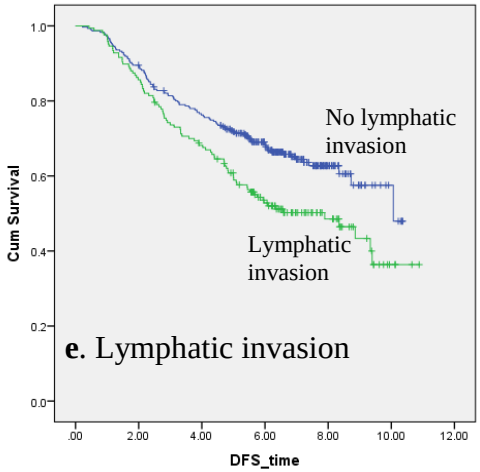
Appendix E: Kaplan-Meier survival plots for the baseline variables associated with disease-free survival

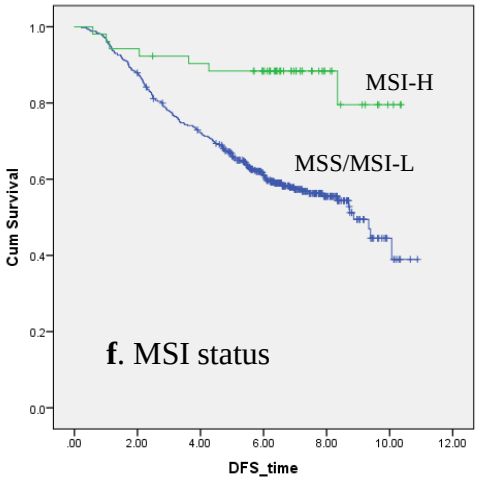







Cum Survival
1.0
0.8
0.6
0.4
0.2
0.0
.00
2.00
4.00
6.00
8.00
10.00
12.00
DFS_time
Colon
Rectum
d. Location


Cum Survival
1.0
0.8
0.6
0.4
0.2
0.0
.00
2.00
4.00
6.00
8.00
10.00
12.00
DFS_time
No lymphatic invasion
Lymphatic invasion
e. Lymphatic invasion


Cum Survival
1.0
0.8
0.6
0.4
0.2
0.0
.00
2.00
4.00
6.00
8.00
10.00
12.00
DFS_time
MSI-H
MSS/MSI-L
f. MSI status

Appendix F: Baseline multivariable model for overall survival

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Sex (male vs female)	0.172	1.274	0.900	1.805
Stage	**<0.001**			
Stage II vs stage I	0.111	1.614	0.896	2.906
Stage III vs stage I	**0.012**	2.141	1.179	3.889
Stage IV vs stage I	**<0.001**	9.560	5.018	18.212
Vascular invasion (+ vs -)	0.372	1.168	0.830	1.643
Location (rectum vs colon)	0.225	1.239	0.877	1.750
MSI status (MSI-H vs MSS/MSI-L)	**0.003**	0.218	0.080	0.597

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable, (+) means present, (-) means absent.

Appendix G: Baseline multivariable model for disease-free survival

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Sex (male vs female)	0.22	1.221	0.887	1.68
Stage	**<0.001**			
Stage II vs stage I	0.111	1.522	0.908	2.551
Stage III vs stage I	**0.005**	2.103	1.245	3.555
Stage IV vs stage I	**<0.001**	5.648	3.152	10.12
Vascular invasion (+ vs -)	0.425	1.138	0.828	1.564
Location (rectum vs colon)	0.088	1.314	0.96	1.797
MSI status (MSI-H vs MSS/MSI-L)	**0.006**	0.339	0.156	0.733

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable, (+) means present, (-) means absent.

Appendix H: Multivariable analysis result for the *MMP3* haplotype (disease-free survival) (recessive genetic model)

a) Adjusting for stage, and MSI status (n =482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.126	1.488	0.895	2.474
Stage III vs I	**<0.001**	2.341	1.424	3.847
Stage IV vs I	**<0.001**	5.641	3.287	9.681
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.293	0.137	0.626
**MMP3* haplotype	0.098	0.712	0.477	1.065

b) Adjusting for stage, age at diagnosis, and MSI status (n =482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	0.265	1.009	0.993	1.025
Stage	**<0.001**			
Stage II vs I	0.132	1.478	0.888	2.457
Stage III vs I	**<0.001**	2.369	1.441	3.896
Stage IV vs I	**<0.001**	5.828	3.386	10.031
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.299	0.140	0.640
**MMP3* haplotype	0.101	0.714	0.478	1.067

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable
*patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes

Appendix I: Multivariable analysis result for the *MMP8* haplotype (disease-free survival) (recessive genetic model)

a) Adjusting for stage, and MSI status (n=482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.133	1.476	0.888	2.455
Stage III vs I	**<0.001**	2.359	1.436	3.876
Stage IV vs I	**<0.001**	5.514	3.204	9.489
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.292	0.137	0.626
**MMP8* haplotype	0.066	1.364	0.979	1.900

b) Adjusting for age at diagnosis, stage, and MSI status (n=482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	0.268	1.009	0.993	1.025
Stage	**<0.001**			
Stage II vs I	0.140	1.467	0.882	2.441
Stage III vs I	**<0.001**	2.387	1.452	3.923
Stage IV vs I	**<0.001**	5.685	3.294	9.811
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.298	0.139	0.637
**MMP8* haplotype	0.069	1.361	0.977	1.895

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable

* patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes

Appendix J: Multivariable analysis result for the *MMP21* haplotype (disease-free survival) (recessive genetic model)

a) Adjusting for stage and MSI status (n=482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.117	1.502	0.903	2.497
Stage III vs I	**<0.001**	2.343	1.426	3.850
Stage IV vs I	**<0.001**	5.845	3.409	10.022
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.303	0.142	0.650
**MMP21* haplotype	0.098	0.719	0.486	1.062

b) Adjusting for age at diagnosis, stage and MSI status (n=482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	0.249	1.009	0.994	1.025
Stage	**<0.001**			
Stage II vs I	0.123	1.492	0.897	2.480
Stage III vs I	**<0.001**	2.370	1.442	3.896
Stage IV vs I	**<0.001**	6.034	3.510	10.373
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.309	0.144	0.661
**MMP21* haplotype	0.095	0.717	0.485	1.059

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable
* patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes

Appendix K: Multivariable analysis result for the *MMP27* haplotype (overall survival) (recessive genetic model)

a) Adjusting for stage and MSI status (n=483)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.113	1.597	0.895	2.851
Stage III vs I	**0.003**	2.399	1.360	4.234
Stage IV vs I	**<0.001**	9.636	5.279	17.591
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.189	0.070	0.511
**MMP27* haplotype	0.107	1.382	0.932	2.050

b) Adjusting for age at diagnosis, stage and MSI status (n=483)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	0.017	1.022	1.004	1.040
Stage	**<0.001**			
Stage II vs I	0.122	1.580	0.885	2.820
Stage III vs I	**<0.001**	2.515	1.423	4.444
Stage IV vs I	**<0.001**	10.616	5.778	19.502
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.191	0.070	0.518
**MMP27* haplotype	0.127	1.360	0.917	2.017

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable
* patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes

Appendix L: Multivariable analysis result for the *MMP27* haplotype (disease-free survival) (recessive genetic model)

a) Adjusting for stage, and MSI status (n=482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.122	1.494	0.899	2.483
Stage III vs I	**<0.001**	2.387	1.453	3.921
Stage IV vs I	**<0.001**	5.560	3.235	9.558
MSI status (MSI-H vs MSS/MSI-L)	**<0.001**	0.292	0.136	0.624
MMP27 haplotype	0.059	1.413	0.986	2.025

b) Adjusting for stage, age at diagnosis, and MSI status (n=482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	0.287	1.009	.993	1.025
Stage	**<0.001**			
Stage II vs I	0.128	1.483	0.892	2.466
Stage III vs I	**<0.001**	2.412	1.468	3.965
Stage IV vs I	**<0.001**	5.727	3.322	9.873
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.297	0.139	0.636
MMP27 haplotype	0.066	1.402	0.978	2.010

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable
* patients homozygous for the most common haplotype vs patients heterozygous for the most common haplotype or patients with the other haplotypes

Appendix M: Multivariable analysis result for the *MMP25* haplotype (overall survival) (co-dominant genetic model)

a) Adjusting for stage, and MSI status (n=483)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.122	1.581	0.885	2.822
Stage III vs I	**0.003**	2.334	1.322	4.121
Stage IV vs I	**<0.001**	9.865	5.409	17.992
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.203	0.075	0.553
MMP25 haplotype coding	0.194			
**MMP25* haplotype	0.108	1.304	0.943	1.803
***MMP25* haplotype	0.742	0.905	0.498	1.642

b) Adjusting for age at diagnosis, stage, and MSI status (n=483)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	0.013	1.023	1.005	1.041
Stage	**<0.001**			
Stage II vs I	0.141	1.546	0.865	2.763
Stage III vs I	**0.002**	2.428	1.373	4.293
Stage IV vs I	**<0.001**	10.883	5.935	19.956
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.207	0.076	0.564
MMP25 haplotype	0.166			
**MMP25* haplotype	0.088	1.326	0.958	1.833
***MMP25* haplotype	0.763	0.912	0.502	1.657

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable
*****patients heterozygous for the most common haplotype vs patients with the other haplotypes; ** patients homozygous for the most common haplotype vs patients with the other haplotypes

Appendix N: Multivariable analysis result for the *MMP25* haplotype (disease-free survival) (co-dominant genetic model)

a) Adjusting for stage, and MSI status (n=482)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Stage	**<0.001**			
Stage II vs I	0.119	1.498	0.901	2.491
Stage III vs I	**<0.001**	2.347	1.428	3.857
Stage IV vs I	**<0.001**	5.804	3.380	9.966
MSI status (MSI-H vs MSS/MSI-L)	**0.002**	0.304	0.142	0.653
MMP25 haplotype coding	0.336			
**MMP25* haplotype	0.243	1.193	0.887	1.606
***MMP25* haplotype	0.574	0.850	0.482	1.499

b) Adjusting for age at diagnosis, stage, and MSI status (n=483)

			95% CI for HR	
Variable	**p-value**	**HR**	**Lower**	**Upper**
Age at diagnosis	0.250	1.009	0.993	1.026
Stage	**<0.001**			
Stage II vs I	0.129	1.483	0.892	2.467
Stage III vs I	**<0.001**	2.368	1.440	3.895
Stage IV vs I	**<0.001**	5.989	3.478	10.313
MSI status (MSI-H vs MSS/MSI-L)	**0.003**	0.310	0.144	0.666
MMP25 haplotype	0.329			
**MMP25* haplotype	0.242	1.194	0.887	1.607
***MMP25* haplotype	0.562	0.846	0.479	1.491

HR: hazard ratio, CI: confidence interval, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable
*****patients heterozygous for the most common haplotype vs patients with the other haplotypes; ** patients homozygous for the most common haplotype vs patients with the other haplotypes

Appendix O: Clinicopathological and treatment-related features for the entire NFCCR cohort

Entire colorectal cancer cohort (n=736)		
Variables	**n**	**%**
Sex		
Female	286	38.9
Male	450	61.1
Location		
Colon	506	68.5
Rectum	230	31.5
Histology		
non-mucinous	644	87.5
Mucinous	92	12.5
Stage		
I	112	15.2
II	244	33.2
III	227	30.8
IV	153	20.8
Grade		
well/moderately differentiated	651	88.4
poorly differentiated	73	10
Unknown	12	1.6
Vascular invasion		
-	398	54.1
+	282	38.3
Unknown	56	7.6
Lymphatic invasion		
-	389	52.9
+	285	38.7
Unknown	62	8.4
Familial risk		
Low	354	48.1
High/intermediate	361	49
Unknown	21	2.8
MSI status		

MSI-L/MSS	634	86.1
MSI-H	73	10
Unknown	29	3.9
BRAF1 mutation status		
-	589	80
+	80	10.9
Unknown	67	9.1
OS status		
Alive	380	51.6
Dead	355	48.2
Unknown	1	0.2
Median OS follow-up: 5.6 years (range: 0.04-11.12)		
DFS status		
no recurrence/metastasis/death	348	47.2
recurrence/metastasis/death	387	52.6
Unknown	1	0.2
Median DFS follow-up: 5 years (range: 0.04-11.12)		
Age		
Median Age: 62.3 years (range: 20.7-75)		

OS: Overall Survival, DFS: Disease Free Survival, MSI-H: microsatellite instability-high, MSI-L: microsatellite instability-low, and MSS: microsatellite stable, (+) means present, (-) means absent.

Appendix P: Chi-square test result for the NFCCR and the study cohorts (stage)

Cohort 505=1, 736=2 Stage Cross tabulation							
			Stage				
			1	2	3	4	Total
cohort 505=1, 736=2)	1	Count	93	196	166	50	505
		Expected Count	83.4	179	159.9	82.6	505
		% within cohort 505=1, 736=2)	**18.40%**	**38.80%**	**32.90%**	**9.90%**	100.00%
		% within Stage	45.40%	44.50%	42.20%	24.60%	40.70%
		% of Total	7.50%	15.80%	13.40%	4.00%	40.70%
	2	Count	112	244	227	153	736
		Expected Count	121.6	261	233.1	120.4	736
		% within cohort 505=1, 736=2)	**15.20%**	**33.20%**	**30.80%**	**20.80%**	100.00%
		% within Stage	54.60%	55.50%	57.80%	75.40%	59.30%
		% of Total	9.00%	19.70%	18.30%	12.30%	59.30%
Total		Count	205	440	393	203	1241
		Expected Count	205	440	393	203	1241
		% within cohort 505=1, 736=2)	16.50%	35.50%	31.70%	16.40%	100.00%
		% within Stage	100.00%	100.00%	100.00%	100.00%	100.00%
		% of Total	16.50%	35.50%	31.70%	16.40%	100.00%

	Chi-Square Tests		
	Value	**df**	**p-value**
Pearson Chi-Square	26.652[a]	3	**0.000**
Likelihood Ratio	28.039	3	**0.000**
Linear-by-Linear Association	17.367	1	**0.000**
N of Valid Cases	1241		

Appendix Q: Chi-square test result for the NFCCR and the study cohorts (vascular invasion)

Cohort 505=1, 736=2 Vascular invasion Cross tabulation					
			Vascular invasion		
			0	1	Total
cohort 505=1, 736=2)	1	Count	308	159	467
		Expected Count	287.4	179.6	467
		% within cohort 505=1, 736=2)	**66.00%**	**34.00%**	100.00%
		% within Vascular invasion	43.60%	36.10%	40.70%
		% of Total	26.90%	13.90%	40.70%
	2	Count	398	282	680
		Expected Count	418.6	261.4	680
		% within cohort 505=1, 736=2)	**58.50%**	**41.50%**	100.00%
		% within Vascular invasion	56.40%	63.90%	59.30%
		% of Total	34.70%	24.60%	59.30%
Total		Count	706	441	1147
		Expected Count	706	441	1147
		% within cohort 505=1, 736=2)	61.60%	38.40%	100.00%
		% within Vascular invasion	100.00%	100.00%	100.00%
		% of Total	61.60%	38.40%	100.00%

Chi-Square Tests

	Value	df	Asymp. P-value (2-sided)	Exact p-value (2-sided)	Exact p-value (1-sided)
Pearson Chi-Square	6.447	1	**0.011**		
Continuity Correction	6.137	1	0.013		
Likelihood Ratio	6.485	1	0.011		
Fisher's Exact Test				0.011	0.007
Linear-by-Linear Association	6.441	1	0.011		
N of Valid Cases	1147				

Pearson Chi-Square p-values are bolded.

Appendix R: Chi-square test result for the NFCCR and the study cohorts (lymphatic invasion)

cohort 505=1, 736=2 Lymphatic invasion Cross tabulation					
			Lymphatic invasion		Total
			0	1	
cohort 505=1, 736=2)	1	Count	298	167	465
		Expected Count	280.5	184.5	465
		% within cohort 505=1, 736=2)	**64.10%**	**35.90%**	100.00%
		% within Lymphatic invasion	43.40%	36.90%	40.80%
		% of Total	26.20%	14.70%	40.80%
	2	Count	389	285	674
		Expected Count	406.5	267.5	674
		% within cohort 505=1, 736=2)	**57.70%**	**42.30%**	100.00%
		% within Lymphatic invasion	56.60%	63.10%	59.20%
		% of Total	34.20%	25.00%	59.20%
Total		Count	687	452	1139
		Expected Count	687	452	1139
		% within cohort 505=1, 736=2)	60.30%	39.70%	100.00%
		% within Lymphatic invasion	100.00%	100.00%	100.00%
		% of Total	60.30%	39.70%	100.00%

Chi-Square Tests

	Value	df	Asymp. P-value (2-sided)	Exact p-value (2-sided)	Exact p-value (1-sided)
Pearson Chi-Square	4.666	1	**0.031**		
Continuity Correction	4.404	1	0.036		
Likelihood Ratio	4.686	1	0.03		
Fisher's Exact Test				0.031	0.018
Linear-by-Linear Association	4.662	1	0.031		
N of Valid Cases	1139				

Pearson Chi-Square p-values are bolded.

Appendix S: Non-parametric Mann-Whitney U-test results comparing median age, overall survival and disease-free survival times for the NFCCR and the study cohort

Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The medians of Age_at_diagnosis are the same across categories of cohort 505=1, 736=2).	Independent-Samples Median Test	.278	Retain the null hypothesis.
2	The distribution of Age_at_diagnosis is the same across categories of cohort 505=1, 736=2).	Independent-Samples Mann-Whitney U Test	.413	Retain the null hypothesis.
3	The medians of OS_time_2010 are the same across categories of cohort 505=1, 736=2).	Independent-Samples Median Test	.000	Reject the null hypothesis.
4	The distribution of OS_time_2010 is the same across categories of cohort 505=1, 736=2).	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.
5	The medians of DFS_time are the same across categories of cohort 505=1, 736=2).	Independent-Samples Median Test	.000	Reject the null hypothesis.
6	The distribution of DFS_time is the same across categories of cohort 505=1, 736=2).	Independent-Samples Mann-Whitney U Test	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Appendix T: Spearman correlation test results for nine of the SNPs in the LD blocks 14 and 15 (Figure 4.1)

Correlations

			MMP27_rs11225388_A_G	MMP27_rs2846707_G_A	MMP27_rs1939015_A_G	MMP27_rs12099177_G_A	MMP27_rs11225389_C_A	rs12418360_C	MMP8_rs12365082_T_A	MMP8_rs7934972_G_A	MMP8_rs12284255_C_A
Spearman's rho	MMP27_rs11225388_A_G	Correlation Coefficient	1	.796**	-.291**	-.166**	.999**	0.014	.997**	-.178**	-.179**
		Sig. (2-tailed)		0	0	0	0	0.758	0	0	0
		N	505	505	505	505	505	505	505	505	504
	MMP27_rs2846707_G_A	Correlation Coefficient	.796**	1	.143**	-.186**	.795**	0.015	.794**	-.236**	-.237**
		Sig. (2-tailed)	0		0.001	0	0	0.741	0	0	0
		N	505	505	505	505	505	505	505	505	504
	MMP27_rs1939015_A_G	Correlation Coefficient	-.291**	.143**	1	-0.066	-.292**	-0.024	-.289**	.621**	.621**
		Sig. (2-tailed)	0	0.001		0.138	0	0.597	0	0	0
		N	505	505	505	505	505	505	505	505	504
	MMP27_rs12099177_G_A	Correlation Coefficient	-.166**	-.186**	-0.066	1	-.166**	-0.009	-.164**	-0.039	-0.04
		Sig. (2-tailed)	0	0	0.138		0	0.846	0	0.379	0.376

	N	505	505	505	505	505	505	505	505	504
MMP27_rs11225389_C_A	Correlation Coefficient	.999**	.795**	-.292**	-.166**	1	0.019	.996**	-.178**	-.179**
	Sig. (2-tailed)	0	0	0	0		0.678	0	0	0
	N	505	505	505	505	505	505	505	505	504
rs12418360_C	Correlation Coefficient	0.014	0.015	-0.024	-0.009	0.019	1	0.015	-0.026	-0.027
	Sig. (2-tailed)	0.758	0.741	0.597	0.846	0.678		0.736	0.554	0.55
	N	505	505	505	505	505	505	505	505	504
MMP8_rs12365082_T_A	Correlation Coefficient	.997**	.794**	-.289**	-.164**	.996**	0.015	1	-.176**	-.177**
	Sig. (2-tailed)	0	0	0	0	0	0.736		0	0
	N	505	505	505	505	505	505	505	505	504
MMP8_rs7934972_G_A	Correlation Coefficient	-.178**	-.236**	.621**	-0.039	-.178**	-0.026	-.176**	1	1.000**
	Sig. (2-tailed)	0	0	0	0.379	0	0.554	0		
	N	505	505	505	505	505	505	505	505	504
MMP8_rs12284255_C_A	Correlation Coefficient	-.179**	-.237**	.621**	-0.04	-.179**	-0.027	-.177**	1.000**	1

	Sig. (2-tailed)	0	0	0	0.376	0	0.55	0		
	N	504	504	504	504	504	504	504	504	504

Green highlight = the correlation of the SNP genotypes are extremely high (r_s >0.99), Red font = SNP genotypes are highly correlated (r_s = 0.8)