Ann-Therese Ali

Targeted Next-Generation sequencing identified novel gene variants involved in hereditary Colorectal Cancer

Master thesis in Molecular Medicine

Trondheim, June 2015

Supervisor: Wenche Sjursen

Norwegian University of Science and Technology

Faculty of Medicine

Department of Medical Genetics

NTNU – Trondheim
Norwegian University of
Science and Technology

Abstract

Colorectal cancer (CRC) is one of the most common types of cancer both worldwide and in Norway. Risk factors and mechanisms contributing to the disease are among dietary and lifestyle and somatic and inherited mutations. CRC is divided in three groups 1. Sporadic CRC where the patients have no family history and no identifiable mutations; 2. Familial CRC where the majority of genetics are unknown but the patients have at least one blood relative, but no specific germline mutation or clear inheritance pattern; 3. Hereditary CRC syndromes where the patients have inherited a single gene mutation in highly penetrant cancer susceptibility genes. The genes known to date to predispose to colorectal cancer are APC, BMPR1A, POLE, SMAD4, the MMR genes among others and these genes are related to the hereditary CRC syndromes. There are other genes which have been found in Genome-wide association studies (GWAS), exome studies or with next-generation sequencing (NGS) to be associated with CRC such as KLLN, AKT1, PIK3CA, OGG1, KIF23 among others. 123 genes some known to be involved in hereditary CRC syndromes and some associated with CRC were sequenced in 95 patients using Haloplex targeted NGS. The purpose for this master thesis was to identify pathogenic variants in these genes that could be of help to explain the increased CRC risk in these patients.

The results from the NGS identified 1268 unique variants which were filtered with the downstream analysis tool FILTUS. After the variants found in only one or two patients were selected. 64 unique variants were left to be further evaluated using the prediction software Alamut. From the 64 variants 25 variants were selected to further investigate because those were found to have most prominent effects on the proteins. Four out of the 25 variants were found to be involved in predisposition to hereditary CRC syndromes; two variants identified in POLE, a variant in BMPR1A and a variant in PTEN. The other variants identified may be involved in CRC predisposition, but further functional studies are needed to determine their function in CRC involvement. There were identified a few false positive variants during the use of Haloplex targeted NGS, but because the rate of these variants were not high this method seems to be a reliable method to use in cancer research.

Acknowledgements

This master thesis was carried out at the Norwegian University of Science and Technology (NTNU) in Trondheim at the Department of Medical Genetics at St. Olavs Hospital.

First of all I would like to express my gratitude to my supervisor Wenche Sjursen for giving me the opportunity to work on this interesting project. Thank you so much for your help, feedback and encouragement during the writing process. Your enthusiasm and passion for the field of genetics is inspiring and contagious.

Many thanks to Maren F. Hansen for help with the laboratory work, interpretation of result data and for reviewing my thesis. Thank you for always being available and for answering all my questions. Want to wish you good luck with your doctoral study.

I would also like to thank Sten Even Erlandsen for taking the time to help with the Bioanalyzer measurements and for sequencing the Haloplex library.

Thanks to Jostein Johansen for processing the data from the next-generation sequencing making the data easier to work with. I appreciate that you took the time to answer my questions.

Thank you also to Liss Anne S. Lavik for showing me around the laboratory and giving me the necessary guidelines to work in the laboratory.

I deeply appreciate working together with Karoline B. Rypdal on some parts of this project. It was very nice to have someone to work with during the laboratory part of the project and to discuss the work that had to be done during the project.

My special thanks goes out to my father, mother and brother for being supportive and for believing in me all the way. I am truly grateful for you always being there for me.

Last but not least I would like to thank my friends, especially Maja Helland for the nice lunches at the Knowledge Center in between writing.

Trondheim, June 2015

Ann-Therese Ali

List of Abbreviations

AC	Amsterdam Criteria
BMP	Bone Morphogenic protein
Bp	Base pairs
CIN	Chromosomal Instability
CRC	Colorectal Cancer
CS	Cowdens Syndrome
EMT	Epithelial-transition mesenchymal
FAP	Familial Adenomateous Polyposis
FCC	Familial Colorectal Cancer
gDNA	Genomic DNA
GI	Gastrointestinal
HDGC	Hereditary Diffuse Gastric Cancer
HMPS	Hereditary Mixed Polyposis Syndrome
HPS	Hamartomatous Polyposis Syndromes
JPS	Juvenile Polyposis Syndrome
LOH	Loss of Heterozygosity
MAP	MUTYH-Associated Polyposis
MCR	Mutation Clustered Regions
MMR	Mismatch Repair
MSI	Microsatellite Instability
MSS	Microsatellite Stable

NSCLC	Non-Small Cell Lung Cancer
OCCS	Oligodontia-Colorectal Cancer Syndrome
PHTS	PTEN Hamartomatous Tumor Syndrome
PJS	Peutz-Jeghers Syndrome
PPAP	Polymerase Proofreading-Associated Polyposis
NGS	Next-Generation Sequencing
TSGs	Tumor-Suppressor Genes

List of Abbreviations for genes in this project

APC	Adenomatous Polyposis Coli
AKR1C4	Aldo – Keto Reductase Superfamily
AKT1	V – Akt Murine Thymoma Viral Oncogene Homolog
AXIN2	Axis Inhibitor 2
BMPR1A	Bone Morphogenetic Protein Receptor 1A
BUB1	BUB1 mitotic checkpoint serine/threonine kinase
BUB1B	BUB1 mitotic checkpoint serine/threonine kinase B
CCDC18	Coiled-Coil Domain Containing 18
CDH1	Cadherin 1, type 1, E-cadherin
CENPE	Centromere Protein E, 312 kDa
CTNNA1	Catenin (Cadherin-associated protein), Alpha 1, 102kDa
DCC	DCC netrin 1 receptor
ENG	Endoglin
EPHB2	Ephrin Receptor B2
FAM166A	Family with sequence similarity 166, member A
FANCM	Fanconi Anemia, Complementation group M
GALNT12	Polypeptide N – Acetylgalactosaminyltransferase 12
GREM1	Gremlin 1, DAN family BMP antagonist
KIF23	Kinesin Family Member 23
KLLN	Killin, p53-regulated DNA Replication Inhibitor
LAMB4	Laminin, Beta 4

LAMC1	Laminin, gamma 1
MAML3	Mastermind-like 3
MYH11	Myosin Heavy Chain 11
MUTYH	MutY Homolog
MRPL3	Mitochondrial Ribosomal Protein L3
NOTCH3	NOTCH 3
NUDT7	Nucleoside Diphosphate Linked Moiety X – Type Motif 7
OGG1	8-Oxoguanine DNA Glycosylase
PIK3CA	Phosphatidylinositor – 4,5 – bisphosphate 3-Kinase, Catalytic Subunit Alpha
POLE	Polymerase DNA directed Epsilon, catalytic subunit
POLD1	Polymerase DNA directed Delta 1, catalytic subunit
PPP1CB	Protein Phosphatase 1, Catalytic subunit, Beta isozyme
PRADC1	Protease – Associated Domain Containing 1
PRSS37	Protease, Serine 37
PSPH	Phosphoserine Phosphatase
PTEN	Phosphatase and Tensin homolog
RAI1	Retinoic Acid Induced 1
SFXN4	Sideroflexin 4
SMAD4	SMAD family member 4
STK11	Serine/Threonine Kinase 11
TBX3	T-box 3
TWSG1	Twisted Gastrulation BMP Signaling Modulator 1

UACA	Uveal Autoantigen with Coiled-Coil Domains and Ankyrin Repeats
ZNF490	Zinc Finger Protein 490

Table of Contents

1. Introduction

1.1 Colorectal cancer

Colorectal cancer (CRC) is a disease that affects the epithelium of the colon and the rectum, and is one of the most common types of cancer both worldwide and in Norway [1, 2]. With more than a million new cases every year, CRC is responsible for about 15% of all the cancers [1, 2]. Risk factors and mechanisms contributing to this disease are among dietary and lifestyle factors and somatic and inherited mutations [3]. There are three major pathways associated with CRC that account for the majority of the CRC cases: Chromosomal instability (CIN), microsatellite instability (MSI), and CpG island methylation phenotype [4].

CRC is divided into three groups: 1. Sporadic CRC which accounts for about 60% of the cases and include patients with no family history and no identifiable inherited gene mutation; 2. Familial CRC (FCC), accounting for about 20-30% of the cases and where the patients have at least one blood relative with CRC or an adenoma, but has no specific germline mutation or clear inheritance pattern [5, 6]; 3. Hereditary CRC syndromes account for approximately 5-10% of the cases where the patients have inherited only a single gene mutation in highly penetrant cancer susceptibility genes. [2, 6] Family history is therefore a big risk factor with a lifetime risk of 10-15% if a first degree relative has CRC, and 30-100% in familial genetic syndromes [7].

The outcome for patients with CRC is dependent on which stage the disease has reached at diagnosis, but the odds for survival normally varies from a 90% 5-year survival rate if the cancers are detected at the localized stage to 10% for individuals that are diagnosed with a distant metastatic cancer [8]. Therefore it is important for at-risk individuals with early detection of CRC due to improved prognosis and a precise understanding of the genetics behind inherited CRC. Early detection of CRC also improves cancer surveillance and prevention strategies, and helps to develop better diagnostic and therapeutic approaches. [5]

1.1.1 Molecular genetics in CRC

The factors behind CRC development are many and they appear to be both complex and heterogenous. Both dietary and lifestyle factors and inherited and somatic mutations contribute to CRC, and the most significant dietary and lifestyle factors seem to be a diet rich with unsaturated fats and red meat, total energy intake, excessive alcohol consumption and reduced physical activity. However, factors that are likely to protect against CRC are nonsteroidal anti-inflammatory drugs, estrogen, calcium and possibly some statins. [3] The process leading to CRC is like any other cancer due to a series of multigene events, and a study by Vogelstein et al. [9] suggested that the progression towards CRC could be due to a series of four genetic events: alteration of APC, K-ras, DCC and p53. The alterations of these genes follow a certain order meaning that the former gene alteration leads to the latter event. [9] A figure of these events can be seen under in figure 1.1.

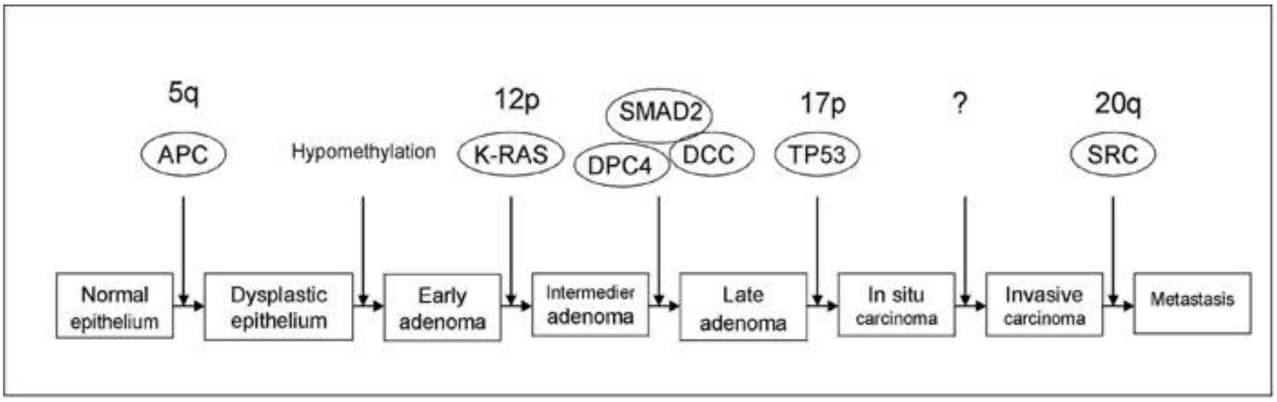


Figure 1.1. Overview of the gene mutations in chronological order that is suggested to lead to CRC, where the mutation in the APC gene initiates the series of mutational events [10].

According to Knudson's two-hit hypothesis mutations in tumor suppressor genes (TSGs) that are inherited will not alone cause tumorigenesis. This is because there is still one healthy copy of the gene in every cell in the body. A cell will not lose its function until the second copy of the gene has turned nonfunctional and Knudson's two-hit hypothesis suggests that this is due to a somatic mutation. [11, 12]

The hereditary CRC cases are as mentioned previously divided in two groups: hereditary CRC syndromes where the genetic cause is known and FCC where the genetics behind the majority of cases are unknown. The most likely cause of FCC is a combination of alterations in high penetrant single genes and low penetrant- and multigenes. Common polymorphisms in genes regulating metabolism or genes regulated by environmental or other genetic factors are examples of this. [5]

Genes known to date to predispose to CRC are APC, BMPR1A, CDH1, POLE, MUTYH, SMAD4, STK11 and the MMR genes among others. These genes are high penetrant and are related to the hereditary CRC syndromes. Other genes such as KLLN, AKT1, PIK3CA, OGG1, KIF23 among others have been found to be associated with CRC in genome wide association studies (GWAS), exome studies or with next-generation sequencing. In GWAS single nucleotide polymorphisms (SNPs) that have moderate or low penetrance are studied to determine their function in disease. The full names of the known genes can be found under a list of abbreviations (see page V).

1.2 Syndromes known to cause hereditary CRC

1.2.1 Lynch Syndrome

Lynch syndrome, also known as Hereditary Nonpolyposis Colorectal Cancer (HNPCC) is responsible for about 2-4% of the hereditary CRC cases, making it the most common cause of hereditary CRC [2]. Lynch syndrome was earlier diagnosed in families that fulfilled the Amsterdam criteria (AC) I which was later modified to AC II so that the extra-colonic cancers could be included. An overview of these criteria can be seen in table 1.1. In the present day it has become clear that not all AC positive families have Lynch syndrome. This was because the cause of MSI was identified which is loss of mismatch repair activity, and this led to the discovery of the genes that cause Lynch syndrome. [4]

Table 1.1. Overview of the Amsterdam Criteria [4].

Amsterdam Criteria I	**Amsterdam Criteria II**
At least three relatives with CRC and the following:	At least three relatives with HNPCC-related cancers (colorectal, endometrial, small bowel, ureter or renal pelvis) and the following:
One should be a first degree relative of the other two	One should be a first degree relative of the other two
At least two consecutive genereations should be affected	At least two consecutive generations should be affected
At least one case of colorectal cancer should be before age 50	At least one case of HNPCC-related cancer should be before age 50
Familial adenomatous polyposis should be excluded in any cases of CRC	Familial adenomatous polyposis should be excluded in any cases of CRC
Verfication of tumors' histopathology	Verification of tumors' histopathology

The syndrome is caused by a germline mutation in one of the MMR genes: MLH1 (MIM #120436), MSH2 (MIM #609309), MSH6 (MIM #600678) and PMS2 (MIM #600259), and has an autosomal dominant inheritance pattern [2]. This syndrome is not described in detail because the MMR genes were not the focus of this study, it was the focus of another master thesis.

1.2.2 Familial Adenomatous Polyposis

FAP is an autosomal dominant syndrome, and it is one of the common inherited CRC syndromes having a prevalence of 1 in 10 000 individuals. The typical traits for classic FAP are development of several up to thousands of colonic adenomas which starts early in adolescence, and will continue to CRC if untreated [2, 5]. Individuals with the classic form of FAP have an average lifespan of about 39 years with untreated CRC, and approximately 95% will have developed CRC at the age of 50. A less severe form of the syndrome is attenuated FAP where there are fewer colonic adenomatous polyps with an average of 30 polyps and the maximum being about a 100. The average age of onset is higher, also individuals with this form of the syndrome will develop polyps and CRC at a later age and the average lifetime risk of CRC is 69%. [5] There are several extra-colonic cancers that take place in FAP such as duodenal cancer, which is the second most common of the extra-colonic cancers in FAP, fundic gland polyps are also common although they do not have a high cancer risk. Gastric adenomas on the other hand have a higher risk towards the development of cancer but they are not all that common.[2] The cause of both classic and attenuated FAP is the germline mutations in the adenomatous polyposis coli (APC; MIM #611731) gene which is a tumor suppressor acting as an antagonist in the WNT signaling pathway [2, 13]. The gene is also involved in processes such as cell migration and adhesion, transcriptional activation and apoptosis. The mutations which are associated with disease have a tendency to cluster in a small region called the mutation cluster region (MCR) and this results in a truncated protein. [13] There have been identified more than 1000 APC variants that cause truncated protein products due to premature stop codons or frameshifts [2].

Classic FAP is diagnosed if at least 100 polyps are identified, whereas attenuated FAP 10 or more but fewer than 100 polyps have to be present. To diagnose attenuated FAP can sometimes be difficult, since the number of polyps can vary with this syndrome and also because it can mimic classic FAP and other syndromes such as MUTYH-associated polyposis (MAP), Lynch syndrome and even sporadic polyp development.[5, 14]

1.2.3 MUTYH-Associated Polyposis

MUTYH-Associated Polyposis (MAP) is an autosomal recessive inherited syndrome where the typical traits are adenomatous polyposis present in the colorectum and an increased risk of CRC [5, 15]. The genetic cause of this syndrome is due to biallelic mutations in the gene MutY homolog (MUTYH; MIM #604933) [5, 16]. MUTYH has the cytogenetic location 1p34.1 and encodes a DNA glycosylase which is part of the base-excision pathway by participating in the oxidative DNA damage repair process [2, 16]. The actual function of MUTYH is to help make sure that G:C to T:A transverions into highly mutagenic bases due to oxidative stress do not occur [5].

The patients usually develop colonic polyposis by the age of 40 even though development of polyps and cancer can take place earlier [5]. Adenomatous polyps dominate in MAP where the patients usually develop an average of about 50 polyps, but there have also been cases with serrated polyps and unlike attenuated FAP hyperplastic polyps are common [2, 5, 15]. There have also been reported cases of MAP that did not show a polyposis phenotype. Extracolonic cancers can also occur in this syndrome where those described are breast-, gastric-, thyroid-, testis- and hematologic cancer. [15]

Criteria for diagnosing MAP have not quite been fully established, but for now the MAP phenotype is considered similar to attenuated FAP. Diagnosing MAP according to genetics will confirm the syndrome and will allow for genetic testing of family members. [5]

1.2.4 Polymerase Proofreading –Associated Polyposis

According to the study by Palles et al. [17] a new hereditary CRC syndrome Polymerase Proofreading-Associated Polyposis (PPAP) has been identified where the cause is germline mutations in Polymerase DNA directed epsilon, catalytic subunit (POLE; MIM #174762) and Polymerase DNA directed delta 1, catalytic subunit (POLD1; MIM #174761) [8, 18, 19]. POLE encodes the catalytic subunit of DNA polymerase epsilon, where the enzyme is involved in DNA repair and chromosomal DNA replication [17, 18]. During the DNA replication POLE is responsible for the synthesis of the leading strand. POLE also has proof-reading capacity through the POLE exonuclease domain which is important for maintenance of replication fidelity. This capacity does not only act on newly misincorporated bases but may also act on mismatches that are produced by non-proof reading polymerases like Polα. [17] POLD1 is also involved in DNA replication and repair where it participates in the mismatch and base excision repair pathways [17, 19]. The gene encodes the catalytic and proof-reading subunit of DNA polymerase delta, which is the equivalent lagging strand polymerase to POLE [17].

PPAP is a dominant inherited syndrome which predisposes to the development of several colorectal adenomas and carcinomas. POLE and POLD1 are both involved in proofreading activity and in patients with this syndrome this proofreading exonuclease activity is impaired due to mutations in these genes. [20]

1.2.5 Hamartomatous Polyposis Syndromes

Hamartomatous Polyposis syndromes (HPS) are a group of rare hereditary genetic autosomal dominant disorders that cover less than 1% of all the hereditary CRCs. HPS includes Peutz-Jeghers syndrome (PJS), Juvenile polyposis syndrome (JPS), PTEN hamartoma tumour syndrome (PHTS) which includes Cowden syndrome (CS) and Bannayan-Riley-Ruvalcaba syndrome (BRRS), and characteristic traits for all these syndromes are the Hamartomatous polyps. These polyps are in themselves benign comprised of cells that are indigenous in the area that they are found in, but these syndromes have a malignant potential to develop both CRC as well as extracolonic cancers.[21]

1.2.5.1 Peutz-Jeghers syndrome

In PJS the hamartomatous polyps occur in the gastrointestinal tract [22]. Characteristics for the PJS polyps are that they are usually multilobulated with a papillary surface with branching bands of smooth muscle covered by hyperplastic glandular mucosa [23]. The consequence of gastrointestinal polyps can be gastrointestinal bleeding, anemia and abdominal pain caused by intussuception, obstruction or infarction [22]. Another characteristic trait with this syndrome is that it causes mucocutanous hyperpigmentation of the lips, buccal mucosa and digits [24]. PJS also has a high rate of extracolonic cancers such as gastric, small bowel, pancreatic, breast, ovarian, lung, cervical and uterine/testicular cancer [2].

To diagnose this syndrome there has been a few criteria proposed: (1) There have to be findings of three or more Peutz-Jeghers (PJ) polyps confirmed histologically; (2) Family history with PJS; (3) Mucocutanous pigmentation that is characteristic and prominent with a family history of PJS; (4) Both mucocutanous characteristic and prominent pigmentation and any number of PJ polyps [25].

PJS can occur due to germline mutations in the serine threonine kinase 11 gene (STK11; MIM #602216) [22]. These germline mutations have been documented in about 70-80% of the patients with PJS where about 15% of the cases part or all of STK11 had been deleted [24]. The function of this gene is complex and is still being researched, but it seems to be a tumor suppressor gene and has been found to regulate the cell cycle, mediate apoptosis, and cellular polarity including other functions [2, 22, 25].

1.2.5.2 Juvenile Polyposis Syndrome

JPS is characterized by juvenile polyps that usually occur throughout the gastrointestinal tract [2]. These polyps appear as spherical and microscopically they are characterized by overgrowth of an oedematous lamina proparia (mucus membranes or mucosa), with inflammatory cells and cystic glands [23]. This syndrome carries an increased risk of CRC and diagnostic criteria according to the World Health Organization (WHO) require one of the following: (1) having more than five polyps in the colon or rectum; (2) having Juvenile polyps present in the gastrointestinal tract; (3) Patients with juvenile polyps having a family history of JPS. [2]

JPS can occur due to a germline mutation in one of the three genes SMAD4 (MIM #600993), BMPR1A (MIM #601299) and ENG (MIM #131195), all related to transforming growth factor-beta (TGF-beta). Mutations in SMAD4 and BMPR1A are each found in approximately 20% of patients with JPS.[2] The BMPR1A gene is located in the same chromosomal region as the PTEN gene, and there have been reported large deletions in both genes. These patients show a more severe form of the syndrome with onset in early childhood or symptoms of both CS and JPS. [23]

1.2.5.3 Cowden Syndrome

PHTS includes patients that clinically have CS and BRRS. The germline mutation of the phosphotase and tensin homolog PTEN (MIM # 601728) gene can be the cause of both these syndromes. Both CS and BRRS are rare syndromes where BRRS is mostly present in the pediatric population whereas CS is most commonly present in adults. [2]

CS is a disease with variable penetrance where traits such as multiple hamartomatous and neoplastic lesions of the skin, mucous, membranes, thyroid, breast, colon, endometrium and brain can be seen [26]. The mutations that occur in the PTEN gene which are associated with CS, are usually point mutations, smaller deletions or insertions [23]. PTEN is a tumor suppressor gene and in approximately 85% of probands with CS there have been identified germline mutations in this gene [26, 27]. PTEN encodes the protein phosphatidylinositol-3, 4, 5-triphosphate 3-phosphatase containing the two domains, a tensin like domain and a catalytic domain which is similar to the dual specificity protein tyrosine phosphatases. This protein unlike the other protein tyrosine phosphatases, dephosphorylates phosphoinositide substrates and also negatively regulates the intracellular levels of phosphatidylinositol-3,4,5-triphosphate in cells.[28] The protein accomplishes this by antagonizing the phosphatidylinositol-triphosphate kinase (PI3K) signaling pathway through its lipid phosphatase activity which results in the following inhibition of the Akt proto-oncogene [27]. It also functions as a tumor suppressor because it negatively regulates the AKT/PKB signaling pathway [28]. The phosphatase activity of the encoding protein regulates the mitogen-activated protein kinase (MAPK) pathway in a negative manner according to Gu et al. [29]. Inactivation or loss of function of PTEN will in the mentioned signaling pathways, cause increased cell survival and uncontrolled cellular proliferation, followed by neoplasia as seen in many human cancers [27].

1.2.6 Syndromes associated with hereditary CRC

1.2.6.1 Hereditary diffuse gastric cancer

Hereditary diffuse gastric cancer (HDGC) is a cancer syndrome with an autosomal dominant inheritance pattern and is caused by germline mutations in the genes Cadherin 1, type 1, E-cadherin (CDH1; MIM #192090) and Catenin (Cadherin associated protein) alpha 1, 102 kDa (CTNNA1; MIM #116805). CRC has been observed in this syndrome in patients belonging to families positive for CDH1.

1.2.6.2 Oligodontia-Colorectal cancer syndrome

Oligodontia is the genetic explanation for severe tooth agenesis, where the characteristics are congenital lack of six or more permanent teeth. It is a very rare disease and is usually related with some multiorgan syndrome. In the study by Lammi et al. [30] it was found that Oligodontia may have a connection to susceptibility for hereditary CRC. The cause of Oligodontia and predisposition to cancer was found to be a nonsense mutation Arg656Stop in the Axis inhibitor 2 (AXIN2; MIM #604025) gene. [30] Mutations in this gene are associated with CRC with defective mismatch repair. [31]

1.3 Other associations with CRC

There have been several studies that have found genes to be associated with CRC and some of these studies are described here. In a study by Alhopuro et al. [32] MYH11 (MIM #160745) was examined whether not it had a mononucleotide tract in its coding sequence since these tracts are vulnerable to mutations under MSI. The study by Bennett et al. [33] identified a hypermethylation in PTEN that resulted in downregulation of KLLN (MIM #612105) through transcription. In another study executed by Gylfe et al. [34] there were identified 14 truncating germline variants in eleven novel predisposing genes in at least two families with CRC. The genes identified were: AKR1C4 (MIM #600451), CCDC18, MRPL3 (MIM # 607118), NUDT7 (MIM #609231), PRADC1, PRSS37, PSPH (MIM #172480), SFXN4 (MIM #615564), TWSG1 (MIM #605049), UACA (MIM #612516) and ZNF490 [34]. DeRycke et al. [7] found CENPE (MIM #117143) and KIF23 (MIM #605064) to include novel missense variants in the susceptibility for FCC. The studies by Smith et al. [35] and Kim et al. [36] identified variants in the OGG1 (MIM #601982) gene that were associated with CRC. In a study by Kokko et al. [37] four heterozygous missense variants that were previously unreported were identified in EPHB2 (MIM #600997). Two studies by Guda et al. [38] identified two somatic and seven germline mutations in the GALNT12 (MIM #608812) gene that were associated with CRC.

The MYH11 gene produces two splice variants SM1 and SM2, which are distinct in the C-terminal tailpiece. In the study by Alhopuro et al. [32] a mononucleotide repeat of 8 cytosines (C8) was observed in the SM2 isoform, and MYH11 was therefore discovered as a candidate MSI colon cancer gene. Mutations that were found during this study were protein-elongating frameshift mutations found in 55% of the CRC cases that exhibited MSI meaning somatic mutations, and also found in the germline of an individual with PJS. There were also discovered two somatic missense mutations in one microsatellite stable (MSS) CRC. All the mutations led to unregulated molecules that showed constitutive motor activity. [32]

In the study by Bennett et al. [33] the hypermethylation upstream of PTEN were detected in 45 out of 123 patients with CS or Cowden-Like syndrome (CSL). The result of the germline methylation was including downregulation of KLLN also disruption of TP53 activation of KLLN by approximately 30%. The study found that the epigenetic modification accounted for one-third of CS individuals negative for germline PTEN mutation and more than 40% of those with CS who were PTEN mutation negative had germline epigenetic inactivation of the KLLN promoter. [33]

The results from the study by Gylfe et al. [34] showed that out of the eleven genes identified four showed loss of the wild-type allele in at least one tumor and a total of seven events with loss of heterozygosity (LOH) were detected, although none showed loss of the mutant allele. This proposes that complete inactivation of these genes is suitable for tumor development and also that these variants are major candidates for CRC susceptibility. Two of the genes that were of particular interest were UACA and TWSG1. This was due to that three out of 96 familial CRC cases were found to have heterozygous truncating variants in UACA and in TWSG1. [34]

The study by DeRycke et al. [7] found that the missense variants identified in KIF23 and CENPE were rare. The variant found in KIF23 was only observed in the ESP database of European Americans, but the CENPE variant was not seen in any of the public databases. Both of these variants were validated and replicated and both are located in previously reported CRC linkage regions. [7]

In the study by Smith et al. [35] the variant identified in OGG1 was a rare inherited nonsynonymous variant with an over representation in patients suffering from advanced CRC compared to population based control subjects [35]. The variant identified was a Gly308Glu substitution and because Glycine at residue 308 through evolution had been much conserved it was predicted that the Glutamic acid substitution would interfere with function. The results of the study showed infrequently biallelic inherited and somatic OGG1 mutations in carriers of OGG1 Gly308Glu and no associated somatic mutator phenotype was observed. This suggests that the variant may play a role as a low-penetrance allele contributing to colorectal tumorigenesis. [35] In the study by Kim et al. [36] the variant discovered was a R154H polymorphism and it was present in patients with FAP, sporadic CRC and in normal controls. R154H was found to be associated with sporadic CRC patients, but did not segregate with cancer phenotypes. The results from the study also showed that there was low possibility of recessive inheritance of R154H, but this still needs to be elucidated. [36]

The study by Kokko et al. [37] found that two of the variants I361V and R568W in EPHB2 were identified in Finnish CRC patients, and the third variant D861N was identified in a UK patient with hyperplastic polyposis (HPP). The fourth variant R80H was identified in a

Finnish patient with CRC and was also found in 1 of 206 familial CRC patients and in 9 of 281 healthy controls and therefore it is likely that this variant might be a neutral polymorphism. The results altogether suggest that EPHB2 may play a limited role in CRC predisposition and that it plays a bigger role in tumor progression rather than in tumor initiation. [37]

The results from the study of 30 MSS colon cancer cell lines by Guda et al. [38] showed that the two somatic mutations identified in GALNT12 were both found in the primary colon tumors from which the cell lines were established, and absent in the normal colon tissues from the same patients. It was also found that these two mutations were within the GALNT12 catalytic and lectin binding domains. The study proved that the two somatic mutations completely inactivated the enzymatic activity of GALNT12, but the wild type GALNT12 allele was found to be retained and expressed in both tumors with the inactivating mutations. [38] In the other study by Guda et al. [38] which was performed to see whether germline mutations in GALNT12 contributed to the development of colon cancer, six of the seven germline variants identified encoded inactive GALNT12 enzymes. [38]

1.4 Next-Generation Sequencing

Next-generation sequencing (NGS) is a term that refers to a rapid evolving high-throughput technology field, that is capable of producing large numbers of DNA sequences in parallel very efficiently, making it a less costly and time consuming method than the earlier technologies used to sequence parts of the human genome, such as Sanger sequencing and fluorescence-based technologies [39-41]. NGS is a useful tool in cancer studies due its ability to not only sequence whole genomes but also focus on specific genomic regions or specific genes using DNA capturing methods. [40]. The NGS workflow is built up of four phases: sample collection, template generation, sequence reactions and detection and data analysis. The template has to be converted into a library of sequencing reaction templates that includes the common steps fragmentation and step size selection, which serve to break the DNA templates into smaller fragments suitable for sequencing. The template generation enables separation and immobilization of the DNA fragment population, thus making it possible for the downstream sequencing reactions to operate while millions of micro reactions are carried out in parallel on each template. To discover structural variants such as insertions, deletions and translocations sequence coverage of approximately 20x to 30x is required to overcome the uneven read distributions and sequencing errors. Biases can be introduced during all steps of NGS and the best example of this is during the template amplification steps. In these steps mutations can be introduced into clonally amplified DNA templates which subsequently masquerade as sequence variants. [41]

1.4.1 Targeted Sequencing

Targeted sequencing is a technique that is very useful in cancer research due its ability to focus on parts of the human genome. During targeted NGS reactions the sequencing reads are distributed to specific genomic locations which equal to higher sequencing coverage and accurate detection of sequence variants regardless of platform error rates. The regions that are targeted needs to be enriched using variable capture strategies such as hybrid capture, microdroplet PCR, or array capture techniques. [41] Targeted sequencing is also a more time and cost-effective method and the data results are considerably more manageable compared to whole exome sequencing. Target enrichment increases sample preparation, cost and time and brings the field of genomics into smaller laboratories. [42] There is only a small percentage of the human genome’s sequence that is characterized and therefore only limited clinically valuable information can be gained from whole genome sequencing. Therefore target sequencing is a more cost effective option for clinical researchers to screen for mutations that could be relevant in the diagnosis and treatment of disease. Targeted sequencing has been useful in screening panels of disease-related genes and also helping to increase the characterization of genetic contribution to different diseases. Due to targeted sequencing being a cost and time effective method it is possible to use genetic testing in diagnosing diseases with complex genetics. [41]

One of the disadvantages with increased throughput of NGS reactions is the read length. Most available sequencing platforms offer on average shorter read lengths than the Sanger sequencing methods, and this restricts the types of experiments that can be conducted by NGS. For instance shorter read lengths may not map or align back to the reference genome uniquely which results in the repetitive sequences of the genome being unmappable in these types of experiments. Another challenge is sequence alignment for regions where there is high diversity between the reference genome and the sequenced genome as it is in structural variants such as insertions and deletions. These challenges are usually solved through the use of longer read lengths or paired-end/mate-pair approaches. [41] Another downside is that the use of panel-based testing can increase the complexity of result interpretation due to an increase in the number of variants of uncertain significance (VUS) [43].

1.5 Aims for this master thesis:

The genetic cause of CRC is only known in the hereditary CRC syndromes accounting for only 5% of the CRCs. Several studies have indicated that some genes could be associated with CRC. The aim for this master thesis was to use a gen panel of several genes reported to be associated with CRC, in order to find the genetic cause for the patients' increased risk of CRC. 123 genes were sequenced in 95 patients where some are known to be involved in hereditary CRC syndromes and some are associated with CRC development using NGS technologies.

For this project the gene list was divided so that the focus was on either the 101 CRC genes or the 22 MMR genes. The aim for this master thesis was to analyze 101 genes which are not involved in the MMR system but some known to be involved in inherited CRC syndromes and some that have been found to be associated with CRC in GWAS or NGS studies. The patients in this project fulfilled the Amsterdam criteria and/or the revised Bethesda guidelines.

2. Material and Methods

The equipment, kits, buffers, solutions and consumables used during this project are listed in table 2.1, 2.2 and 2.3.

Table 2.1. Overview of the equipment used in this project

Description	Vendor/manufacturer	ID
Agilent 2100 Bioanalyzer	Agilent Technologies	Cat: G2940CA
Automat pipette		
Benchtop microcentrifuge, Galaxy Mini	VMR™ International	Cat: 93000-196
Benchtop rotator FSR20	Grant Boekel	
Biohit Eline Pro (Pipette)		
Biohit Pipette (Multichannel)		
Dynal Invitrogen Bead Separations	Invitrogen	
Eppendorf Centrifuge 5810R	VMR™ International	
Eppendorf vortex mixer PCR 96 Tube	Mixmate®Eppendorf	Cat: 5353000.014
Geneflash Bio Imaging system	SynGene	
Iprep™ purification instrument	Invitrogen	Cat: 10000
KMS1 minishaker vortexer	IKA®	
Magnetic Particle Concentrator	Dynal MPC®, Life Technologies	Batch: 44/55
Microplate Sealer ALPS™ 25	ThermoScientific	
Multichannel pipettes, Finnpippette	ThermoScientific	
Nanodrop® ND-1000 spectrophotometer	ThermoScientific	
Plastic Pipette, Disposable, Sterile	Sterilin	
Qubit 2.0 Fluorometer, Invitrogen™	Life Technologies	Cat: Q32866
Thermal Cycler 2720, Applied Biosystemes®	Life Technologies	Cat: 4359659

Table 2.2. Overview of kits, buffers and solutions

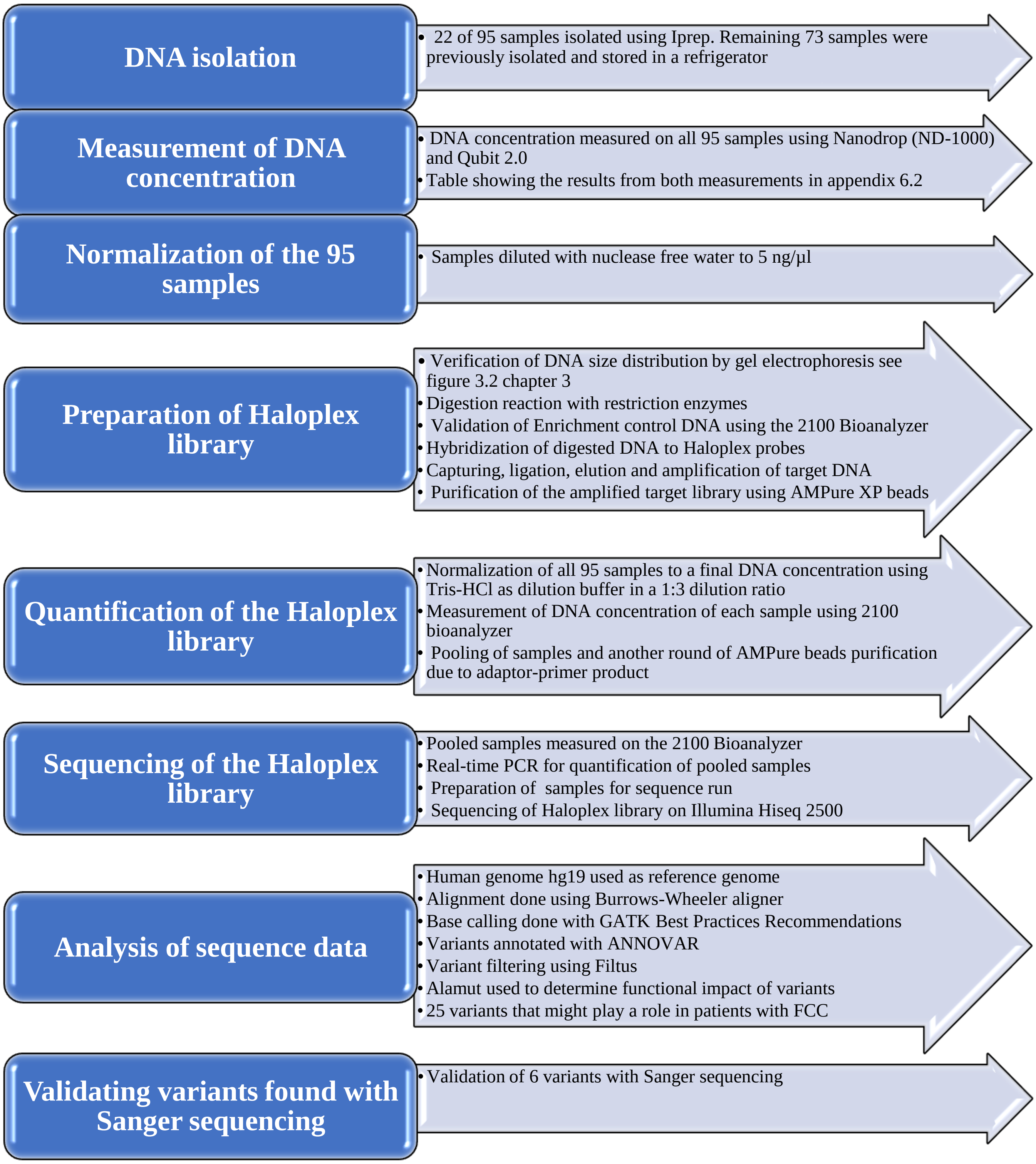
Description	Vendor/manufacturer	ID
Acetic acid solution 2 M		Lot: SLBH 6779V
Elution Buffer (EB) 250 ml	Qiagen Gmblt	Cat: 19086 Lot: 145046057
Haloplex Target Enrichment Kit, 96 reactions	Agilent Technologies	Cat: # 5190-5534 Lot: 0006246792
HCl solution 0.3 M for Nanodrop		
Invitrogen^{TM} Qubit® dsDNA high sensitivity Assay Kit	Life technologies	
Iprep^{TM}Purelink^{TM}gDNA Blood Kit	Invitrogen	Lot: 1603453
NaOH 10 M		Lot: 1168043
Tris-HCl, pH 8.0		
Tris 10 mM for Nanodrop		

Table 2.3. A list of the consumables used in this project

Description	Vendor/manufacturer	ID
Agencourt AMPure® beads	Beckman Coulter Inc	Lot: 14060800
E-Gel iBase^{TM} 2% agarose	Invitrogen	Lot: B16074
Herculase II Fusion DNA Polymerase	Agilent Technologies	Cat: 600677-51 Lot: 0006212697
High Sensitivity DNA Chips	Agilent Technologies	Lot: SF04BK50
High Sensitivity DNA Reagents	Agilent Technologies	Lot: 1420
Ladder 4 DNA Molecular Weight Marker IV (0.07-19.3 Kb)		Lot: 11799634
Nuclease Free Water		

2.1 Workflow of methods used in this project for 123 CRC genes

Exons including splice site regions, 5’- and 3’ UTRs for 123 genes were sequenced using DNA samples from 95 patients. The genes were sequenced on a Illumina HiSeq2500 platform. The figure below lists the methods used in this study.



DNA isolation

- 22 of 95 samples isolated using Iprep. Remaining 73 samples were previously isolated and stored in a refrigerator

Measurement of DNA concentration

- DNA concentration measured on all 95 samples using Nanodrop (ND-1000) and Qubit 2.0
- Table showing the results from both measurements in appendix 6.2

Normalization of the 95 samples

- Samples diluted with nuclease free water to 5 ng/µl

Preparation of Haloplex library

- Verification of DNA size distribution by gel electrophoresis see figure 3.2 chapter 3
- Digestion reaction with restriction enzymes
- Validation of Enrichment control DNA using the 2100 Bioanalyzer
- Hybridization of digested DNA to Haloplex probes
- Capturing, ligation, elution and amplification of target DNA
- Purification of the amplified target library using AMPure XP beads

Quantification of the Haloplex library

- Normalization of all 95 samples to a final DNA concentration using Tris-HCl as dilution buffer in a 1:3 dilution ratio
- Measurement of DNA concentration of each sample using 2100 bioanalyzer
- Pooling of samples and another round of AMPure beads purification due to adaptor-primer product

Sequencing of the Haloplex library

- Pooled samples measured on the 2100 Bioanalyzer
- Real-time PCR for quantification of pooled samples
- Preparation of samples for sequence run
- Sequencing of Haloplex library on Illumina Hiseq 2500

Analysis of sequence data

- Human genome hg19 used as reference genome
- Alignment done using Burrows-Wheeler aligner
- Base calling done with GATK Best Practices Recommendations
- Variants annotated with ANNOVAR
- Variant filtering using Filtus
- Alamut used to determine functional impact of variants
- 25 variants that might play a role in patients with FCC

Validating variants found with Sanger sequencing

- Validation of 6 variants with Sanger sequencing

Figure 2.1. Flowchart of the methods used in this project

2.2 Material and preparation of samples before library preparation

The patient material used for this study was gDNA isolated from EDTA preserved whole blood. Samples from 95 patients that fulfilled the AC and/or the revised Bethesda guidelines (RBG) [44] were chosen for sequencing and these are listed in appendix 6.1. DNA had to be isolated for 22 of the 95 samples using the Iprep^{TM} Purelink^{TM} gDNA blood kit from invitrogen with the Iprep instrument. The manual for the instrument Iprep was used as a procedure for the DNA isolation [45]. Two patients had two blood samples each. Sample 33 and 46 were from one patient and samples 51 and 87 from the other.

The DNA concentration of the samples were measured on both the spectrophotometer Nanodrop (ND-1000) and the spectrofluorometer Qubit 2.0 according to the manufacturer's instructions [46, 47]. The Nanodrop, measured the absorbance of the DNA while Qubit 2.0 measured the fluorescence. The reagents for both Nanodrop and Qubit are listed in table 2.4 and 2.5.

Table 2.4. Reagents, volume and application for Nanodrop concentration measurement of 95 DNA samples

Reagent	Volume	Application
HCl (hydrochloric acid)	2 µl	For washing
H_2O (water)	2 µl	Start-up of the spectrophotometer
Tris buffer	2 µl	Used as a blank

Table 2.5. Reagents and volume for Qubit 2.0 concentration measurement of 95 DNA samples

Solutions	Volume
Total solution in tubes	200 µl
Total sample volume	2 µl
Total buffer + fluorochrome	19,9 µl
Volume in tubes of solution	**198 µl**

Following the DNA concentration measurement, all samples were diluted with nuclease free water to a DNA concentration of 5 ng/µl. A table of the volumes in the dilutions can be found in appendix 6.2.

2.3 Library preparation for NGS according to Agilent Technologies

Before the library preparation, the size distribution of the undiluted DNA samples was verified using gel electrophoresis to see if there had been any smearing below 2.5 kb, which indicates sample degradation. The samples tested were samples 1-10 because these were the oldest samples and also sample 86 were tested since this sample was from year 2014, so it could be compared with the results from samples 1-10. The ladder was diluted with nuclease free water in ratio 1:10. The gel was run for 30 minutes and the results can be seen in a figure 3.2 in chapter 3.

For the library preparation a custom made Haloplex Target enrichment kit for 96 samples was used. The sample preparation was executed according to the HaloPlex Target Enrichment System protocol for Illumina Sequencing [48]. The HaloPlex target enrichment system by Agilent technologies uses a capture method based on hybridization and amplification of the gDNA fragments, where the target DNA is first digested by different restriction enzymes to generate a library of gDNA restriction fragments. The digested DNA is then hybridized to the HaloPlex probes for target enrichment and sample indexing resulting in circularized gDNA fragments where sample indexes and Illumina sequencing motifs have been incorporated. The DNA-probe hybrids which contain biotin allow for capture with streptavidin-coated magnetic beads and DNA ligase is used to close gaps in the circularized DNA- probe hybrids.

The final step is PCR amplification of the targeted fragments so that a sequencing-ready target enrichment sample can be produced. [48] A figure of the HaloPlex target enrichment workflow is shown in figure 2.2 below.

All samples were diluted 1:3 with Tris-HCl and the concentration of each sample library was measured using Agilent 2100 Bioanalyzer. The concentration was used to pool equimolar amounts (10 ng) of each sample. The expected concentration in the final pool and three measurements with Bioanalyzer after pooling the samples can be seen in appendix 6.3.

Before sequencing the library a real-time PCR was performed to quantify the pooled samples. The library was sequenced on the Illumina HiSeq 2500 platform. There were 123 CRC related genes where the exons including splice-site regions and 5’ and 3’ UTRs were sequenced for all the 95 patients. A list of the all the genes that were sequenced due to their involvement in CRC is listed in appendix 6.4.

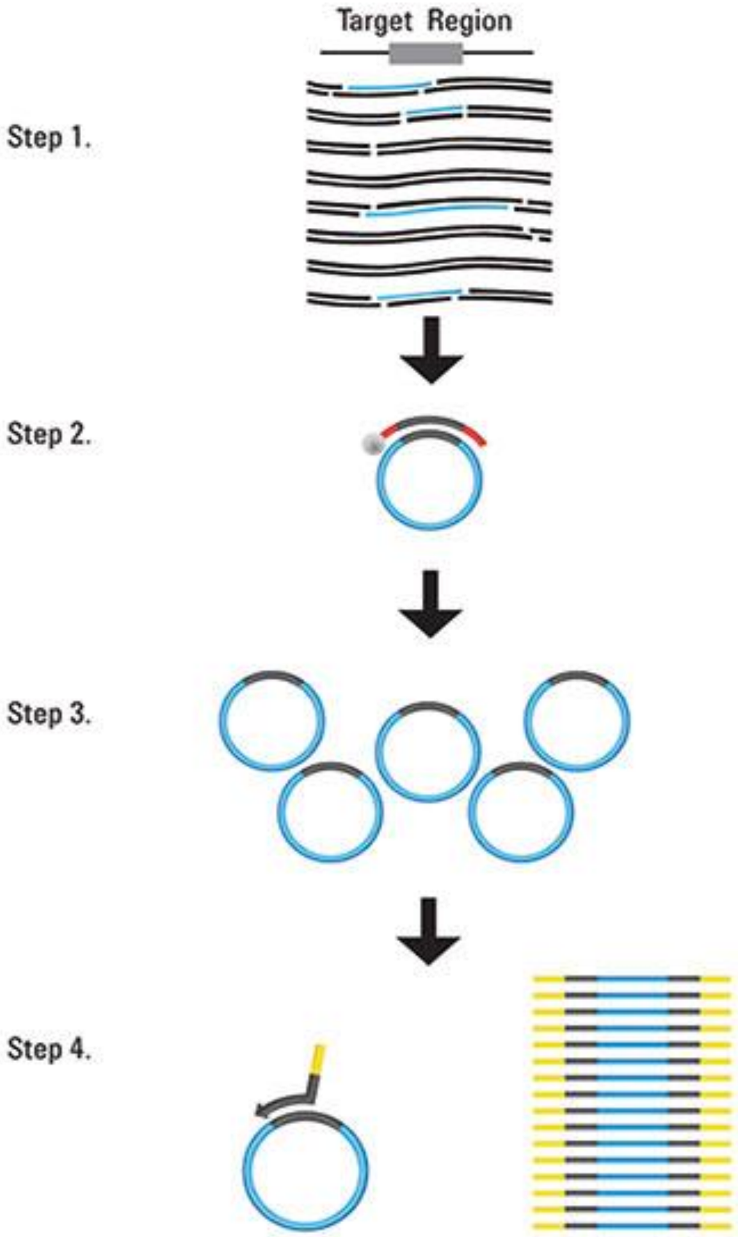


Figure 2.2. Overview of workflow for HaloPlex target enrichment. Step 1: Target gDNA is digested by restriction enzymes into restriction fragments. Step 2: Digested DNA is hybridized to HaloPlex probes for target enrichment and sample indexing, giving circularized DNA fragments. Step 3: Capturing of DNA-probe hybrids with streptavidin-coated magnetic beads. Step 4: Amplification of targeted fragments by PCR producing a target-enriched sample ready to be sequenced. [49]

2.4 Data analysis after NGS

After sequencing the Haloplex library the sequence data was interpreted by chief engineer, Jostein Johansen at BioCore NTNU. The human genome hg19 was used as a reference sequence and the alignment was done using the Burrows-Wheeler-Aligner [50]. In this study the variant calling was done according to GATK Best Practices Recommendations [51, 52] using GATK version 3.1 [53], including local realignment around indels, recalibration of quality scores and quality control of called variants. To analyze the regions targeted from the sequencing experiment, a list of these regions were used as additional information when running GATK to reduce running time of the pipeline. The variants were annotated with a software called ANNOVAR [54].

2.5 Interpreting sequence data

Filtering of variants was done using FILTUS [55] which is a tool for downstream analysis used in high-throughput sequence projects. It was used to filter out variants with a minor allele frequency (MAF) >1% in the 1000 genomes database and all variants present in dbSNP138 to select only rare variants. The synonymous variants and variants of low quality were also removed. The 22 MMR genes and variants with coverage lower than 10 were also excluded. The MMR genes were removed because they were not the focus for this study. An overview of the filters used with details can be seen in table 2.6. The list of the remaining variants after the filtration can be found in appendix 6.5 which also includes information about what effect the mutation has on the protein.

Table 2.6. Filters applied to variants in Filtus

Name of function/database	Filter	Parameters	Keep if missing
Exclude genes	22 MMR genes		
Exonic Func	Not equal to	synonymous SNV	
1000 genomes	Less than	0.01	Ticked off
dbSNP138	Does not contain		Ticked off
FILTER	Equal to	PASS	

2.5.1 Evaluation of variants

After the filtration of variants the software Alamut visual version 2.3 by Interactive biosoftware was used to determine whether or not the mutation had a damaging effect on the protein. The information about the effect the mutation had on the protein was predicted by different prediction programs which are a part of the software such as AlignGVGD, SIFT, MutationTaster and Polyphen-2. [56]

The variants were colour coded according to their effect on the protein and appendix 6.5 shows an overview of these variants. The variants with the most prominent effect on the protein were chosen for further research.

2.6 Validation of variants with Sanger sequencing

The Sanger sequencing was performed by the clinical lab, and only a few variants were validated. The Sanger sequencing was done according to procedures at the clinical lab.

First a PCR was performed to amplify the fragments that were going to be sequenced. Then purification of the PCR-product was done with the reagent A'SAP. This purifying reagent eliminates excess primers and nucleotides enzymatically without eliminating any PCR product.

The sequencing-PCR was then performed which was the process where single stranded DNA was amplified. The sequencing reagents contained fluorescence labeled nucleotides which were attached to the end of each fragment. The sequence reaction produced fragments of different lengths, but the number of each fragment made was random. During the sequence reaction the annealing temperature was specific for each primer used, and it should be similar to the temperature used during the PCR-reaction and no higher than 60˚C. Sequencing was done in both directions to sequence all the way to the opposite primer and also because eventual findings could be double checked.

After the sequence reaction another purifying step is necessary to eliminate excess fluorescence labeled nucleotides that were not incorporated in the actual sequence, salts and other charged molecules that could affect the sequencing performed with capillary electrophoresis. The reagent used for this was BigDye XTerminator Purification kit.

The instrument used for the capillary electrophoresis is the ABI PRISM 3130xl or the ABI PRISM 3730 Genetic Analyzer. The product from the sequence reaction is first injected electro kinetically into the capillaries which are filled with polymer. Then the negatively charged DNA migrates towards the positively charged electrode and close to this electrode the fragments migrate through a laser beam. When the fluorescence comes in contact with the light from the laser a spectrum from each of the four nucleotides is produced. There is also a CCD camera that detects the signals as the fragments in increasing length passes by the detection cell during the electrophoresis. Data collection software is used to convert the fluorescence signals into digital data and then to an electrogram which is processed in other analyze programs such as Seqscape. A figure of an electrogram can be seen below.

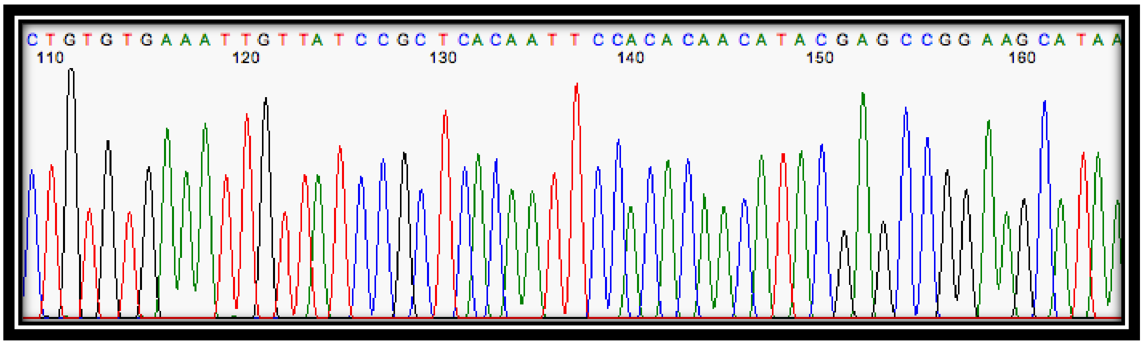


Figure 2.3. The picture shows an electrogram from Sanger sequencing [57]

3. Results

In this study 123 genes related to CRC in 95 patients were sequenced using NGS technology to discover novel variants involved in the development of CRC in high-risk individuals. The focus for this study was 101 CRC related genes. The results were interpreted using the NGS downstream analysis tool FILTUS and the software program Alamut. Some interpreted high risk variants were further investigated.

3.1 Measurement of DNA concentration

DNA concentrations were measured on Nanodrop ND-1000 and Qubit 2.0 to normalize all samples into an equal DNA concentration. The results can be seen in figure 3.1 and a table with the DNA concentration and the amount of DNA necessary to obtain the final DNA concentration of 5 ng/µL for each sample is listed in appendix 6.2.

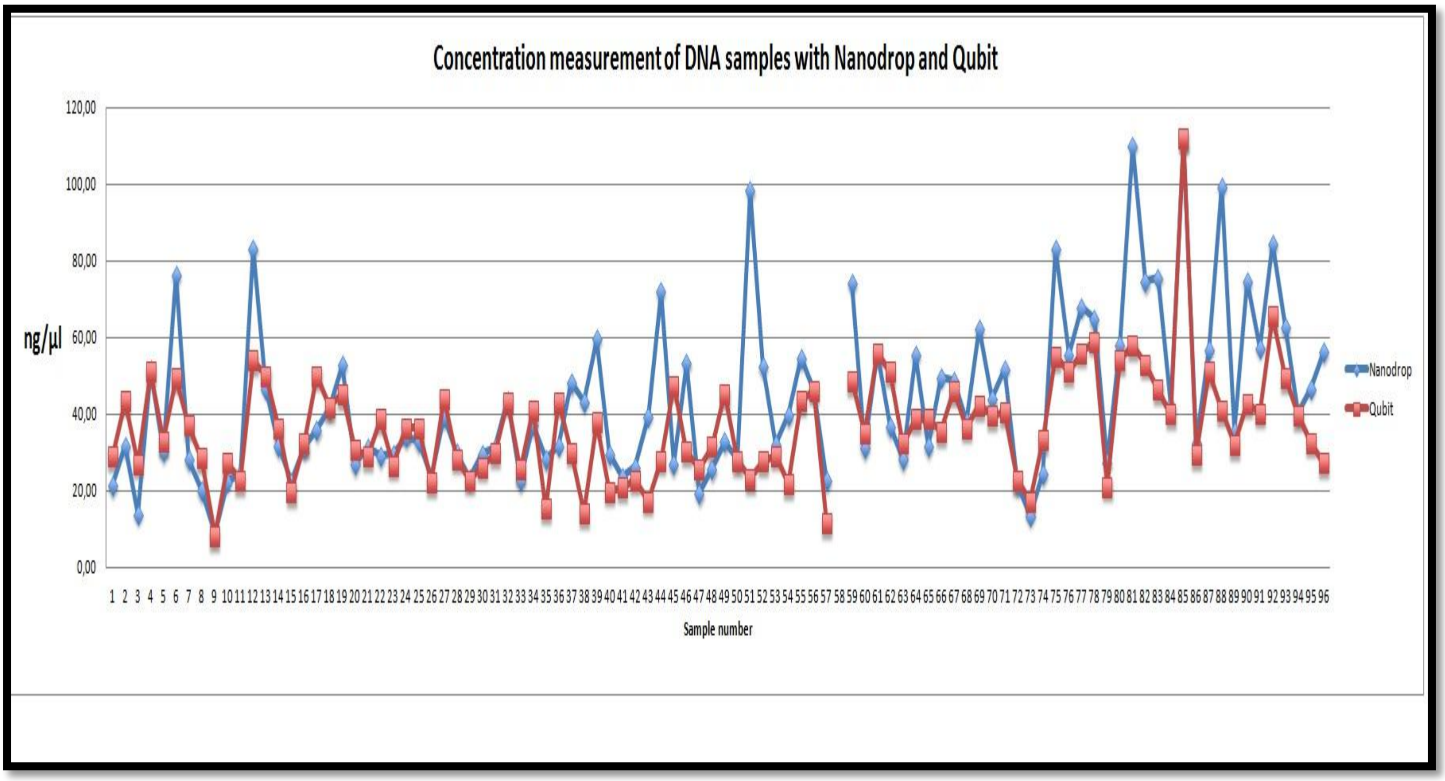


Figure 3.1. A graph displaying DNA concentrations of the 95 samples measured on both Nanodrop and Qubit. The blue line represents results from Nanodrop and the red line Qubit 2.0.

The DNA concentration measurement results showed that the sample concentration measured on Nanodrop was higher in some samples compared to samples measured with Qubit, and therefore the results from the Qubit measurement were used for normalization of the samples. The reason for the elevated concentration could be due to a systematic error in Nanodrop or it could have been because the samples were not mixed properly before measuring.

The size distribution of undiltued DNA samples were verified with gel electrophoresis to see if the DNA in the samples were degraded, and the results can be seen in figure 3.2 below.

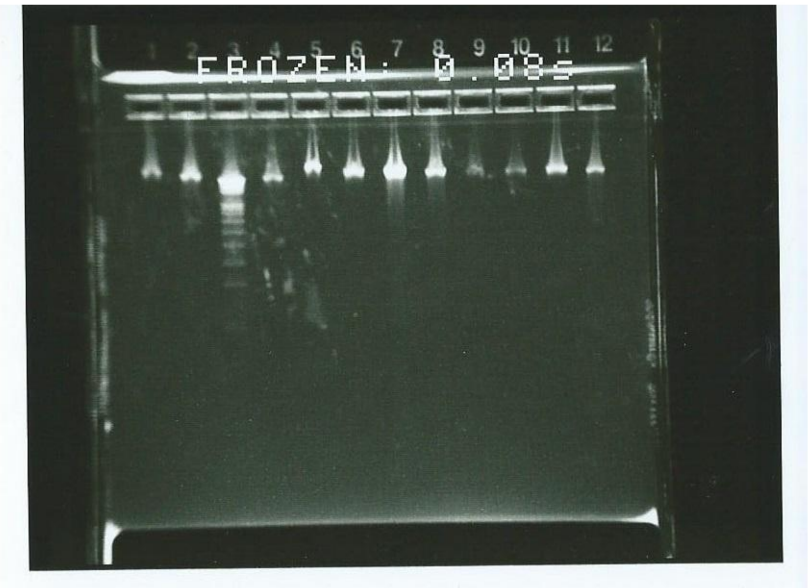


Figure 3.2. The picture shows the results from a gel electrophoresis that tested samples 1-10 and sample 86 to verify the size distribution of DNA. The samples are in chronological order except for sample three which is the ladder meaning that well 4 contains sample nr. 3.

The result from the gel electrophoresis shows that all the fragments were larger than band 9-10 of the ladder well nr 3 in figure 3.2, and these bands were approximately 2.5 kb. This means that there was no smearing below 2.5 kb and therefore no degradation of the DNA in the samples.

3.2 Preparation of Haloplex library

Before the Haloplex library could be pooled and sequenced the enrichment and the quantity of the enriched target DNA had to be validated in each sample. This was done by using the 2100 Bioanalyzer and an example of an electropherogram from one of the samples can be seen in figure 3.3.

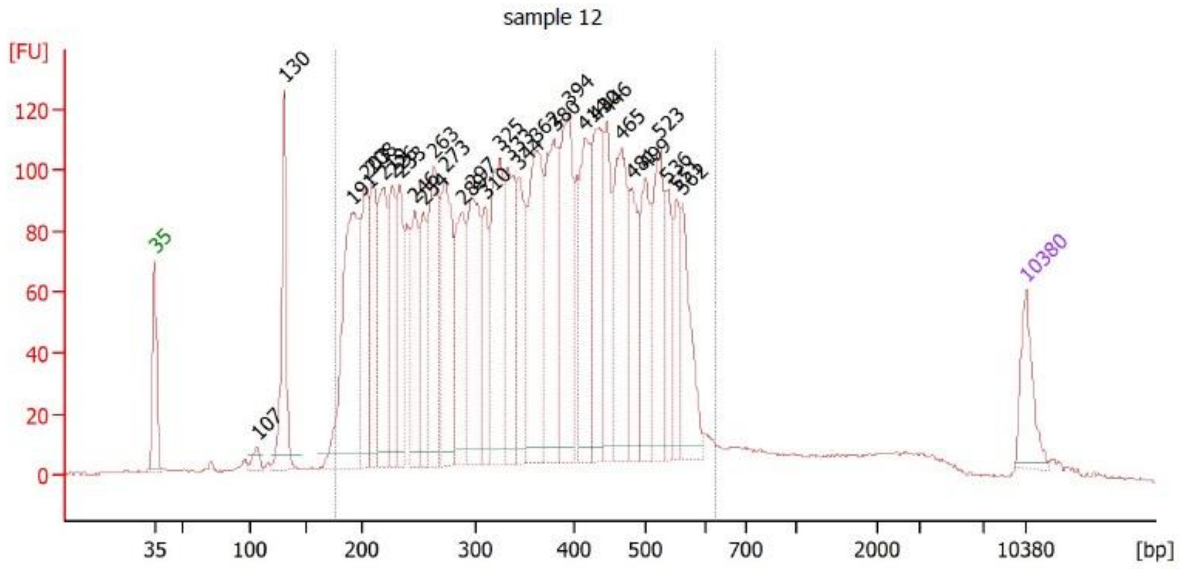


Figure 3.3. Here is an example of an electropherogram from 2100 Bioanalyzer of sample 12 where the largest peak is the sample itself which lies between 200-600 bp. The x-axis displays the number of base pairs whereas the Y-axis plots the intensity of the sample. The peaks with the coloured numbers on top are the lowest and the upper markers of the ladder. The concentration of this sample is 7.66 ng/µL which can be accepted since it is less than 10 ng/uL.

The results from the Bioanalyzer measurement are listed in appendix 6.3 as well as the volume used to obtain equal molar amounts of each sample prior to pooling of the samples. The results showed that the target enrichment process was successful since the Bioanalyzer measurements showed that a library was obtained. The concentration of each sample (appendix 6.3) was acceptable since the concentration was supposed to be below 10 ng/µl. The expected concentration was consistent with actual concentration of the pooled samples as can be seen in the table.

The mean coverage across all samples was 258.18 and the table with mean coverage for each sample is listed in appendix 6.6. The standard deviation of the total coverage is 258.18 ± 57.76. On average, 86.72 % of the target region were covered with >20 reads. The standard deviation of the regions covered with >20 reads is 2162 ± 42.92. The table with regions covered with >20 reads for each sample is listed in appendix 6.7.

3.3 Interpretation of sequence data

The result from the NGS was filtered with the downstream analysis tool FILTUS and the number of variants before any filters were added was 1268 unique variants in 123 genes. A table of the number of total variants in each patient is listed in appendix 6.8. From this table it can be seen that the number of variants in sample 33 and sample 46 both samples from the same patient, are not equal. Before filtering away the variants of low quality there were 305 variants in 80 genes in sample 33 and 338 variants in 77 genes in sample 46. The number of variants after excluding the variants of low quality were 229 variants in 71 genes in sample 33 and 228 variants in 77 genes in sample 46. The variants that differ between the samples are two nonframeshift insertion variants found in the BLM gene c.2318_2319insAGA:pS773delinsRD and c.2319_2320insCGG:pS773delinsSR and one nonframeshift insertion found in the MRPL3 gene c.471_472insTCT:p.A158delinsSA. All three variants were found in sample 46 and not in sample 33. Sample 33 had a nonsynonymous variant in the PSPH gene c.T549:p.D183E that was not found in sample 46. The number of variants in samples 51 and 87 that were also from the same patient were also unequal. Before the variants of low quality were excluded there were 316 variants in 78 genes in sample 51 and 297 variants in 78 genes in sample 87. After the low quality variants were excluded 217 variants in 72 genes remained in sample 51 and 218 variants in 70 genes in sample 87. The variants that differ between these samples are two nonframeshift insertions in BLM and one nonframeshift insertion in MRPL3 all three found in sample 51, which are the same variants found in sample 46.

To make the number of variants more manageable and to focus only on the rare variants, filters were added to exclude variants with a MAF >1% which were present in the databases 1000 genomes and dbSNP138. The other variants that were excluded were the synonymous variants, variants of low quality, the 22 MMR genes and the variants with coverage lower than 10. The table with the details of the filters applied can be found in chapter 2 in table 2.6. After the filtration 1171 variants were excluded leaving 97 unique variants in 54 genes. To further decrease the number of variants the variants found in many patients were excluded due to the fact that they were thought to be more common in the population than those found only in a few patients. The variants that were chosen to look closer into were those found in one or two patients. Information about the predicted consequences of the mutation at protein level was determined using the Alamut software. Appendix 6.5 shows the remaining variants after filtration with information about the predicted effect the mutation has on the protein. As can be seen from the table in appendix 6.5 there were 64 unique variants in 41 genes to do further work with after the selection of variants. A table with number of variants found in each patient after the filtration is listed in appendix 6.9. The variants that are listed in appendix 6.5 have mostly been found in one or two patients, but some variants were chosen that were found in up to ten patients, because these variants were in genes known to cause inherited CRC syndromes. The table shows that there are five types of mutations: 45 missense, 5 frameshift insertions, 5 nonframeshift insertions, 7 frameshift deletions and 2 nonframeshift deletions, and more than half of the mutations are according to Alamut damaging for the protein.

The table in appendix 6.5 was used to select the variants to investigate further to see if these could be involved in the cause of CRC for the high-risk individuals. The variants that were selected were those predicted by Alamut to have a damaging effect on the protein, and these mostly included the frameshift variants but also a few missense variants. The reason why only a few variants were chosen and not all the variants that were according to Alamut damaging to the protein, was because these were thought to be the most interesting variants to further investigate. Usually the frameshift mutations are more damaging than missense mutations because they change the reading frame completely, although this is not always the case. The few missense variants that were decided to investigate further were chosen because some of them are known in the development towards CRC. The number of variants found to investigate further was 25 unique variants.

Some of the variants chosen are found in genes that are known to cause hereditary CRC syndromes and these are listed in table 3.1. The remaining variants are found in genes associated with CRC and they are listed in table 3.2.

Table 3.1. A list of variants in genes known to cause hereditary CRC syndromes. The variants found to be false positive with Sanger sequencing are highlighted in red.

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
POLE	6,29,30	NM_006231:c.1373A>T:p.Y458F	Class C0	Deleterious (score 0)	Disease causing (P-value: 1)	Probably damaging (Humdiv 1.000 + humvar 1.000)
POLE	44	NM_006231:c.824A>T:p.D275V	Class C0	Deleterious (score 0)	Disease causing (P-value:1)	Probably damaging (Humdiv 1.000 + humvar 1.000)
APC	14,24,46,47,49,59-61,72,95	NM_001127511:c.3086_3087insTCGG:p.Lys1030Argfs*2	N/A	N/A	N/A	N/A
BMPR1A	32	NM_004329:c.785T>C:p.V262A	Class C0	Tolerated (score 0.15)	Disease causing (P-value: 1)	Possibly damaging (Humdiv 0.923 + humvar 0.884)
GREM1	30,31,60	NM_013372:c.196_197insT:p.Thr66Ilefs*35	N/A	N/A	N/A	N/A
PTEN	35,48	NM_000314:c.377C>T:p.A126V	Class C0	Tolerated (score 0.29)	Disease causing (P-value: 1)	Probably damaging (Humdiv 1.000 + humvar 0.998)
STK11	60	NM_000455:c.459_460insAGA:p.Ala153_His154insArg	N/A	N/A	N/A	N/A

Table 3.2. Variants in genes associated with CRC

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
AKT1	32	NM_001014431:c.206G>C:p.R69P	Class C35	Deleterious (score 0.01)	Disease causing (P-value: 0.995)	Possibly damaging (Humdiv 0.792 + humvar 0.667)
AKT1	46	NM_001014431:c.520C>T:p.R174C	Class C0	Deleterious (score 0)	Disease causing (P-Value: 1)	Possibly damaging (Humdiv 0.900 + humvar 0.800)
BUB1	49,65	NM_ 004336.4: c. 447_448insTCT p.Glu149_Thr150insSer	N/A	N/A	N/A	N/A
BUB1B	11,45,50,62	NM_001211:c.2252_2253insAGA:p.Pro751_Lys752insAsp	N/A	N/A	N/A	N/A
BUB1B	11,45,62	NM_001211:c2253_2254insCGG:p.Pro751_Lys752insArg	N/A	N/A	N/A	N/A
DCC	46	NM_005215:c.1664_1665insCGAGAT:p.Asn555_Gly556insGluIle	N/A	N/A	N/A	N/A
FAM166A	16,22,52	NM_001001710:c.751_752del:p.Leu251Valfs*2	N/A	N/A	N/A	N/A
FANCM	17,28	NM_020937:c.5607_5608del:p.Glu1870Aspfs*4	N/A	N/A	N/A	N/A
KIF23	71	NM_138555.2 c.610_618del p.Phe204_Lys206del	N/A	N/A	N/A	N/A
LAMB4	78	NM_007356:c.5265delA:p.Lys1755Asnfs*11	N/A	N/A	N/A	N/A

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
LAMC1	14,51,71,79	NM_002293:c.4579_4580del:p.Leu1527Glyfs*7	N/A	N/A	N/A	N/A
MAML3	3,37	NM_018717:c.1513_1514del:p.Gln505Alafs*21	N/A	N/A	N/A	N/A
MAML3	3,18,34,41,52,59,64,77,90	NM_018717:c.1506delG:p.Gln502Hisfs*20	N/A	N/A	N/A	N/A
NOTCH3	34	NM_000435:c.3733_3734insT:p.Thr1245Ilefs*20	N/A	N/A	N/A	N/A
PIK3CA	38	NM_006218:c.107_108insAGAT:p.Cys36fs*	N/A	N/A	N/A	N/A
PPP1CB	72	NM_002709:c.469_470insAGATC:p.Cys157*	N/A	N/A	N/A	N/A
RAI1	35,48	NM_030665:c.838_843del:p.280_281del	N/A	N/A	N/A	N/A
TBX3	66	NM_016569.3.c.1893del p.Asn632Thrfs*257	N/A	N/A	N/A	N/A

3.4 Validation of variants found with Sanger sequencing

The variants validated were in genes included in diagnostic gene testing at the Medical Genetic Laboratory at St. Olavs Hospital. Thus primers for these genes were available, and included the two POLE variants, the PTEN variant, the BMPR1A variant, the STK11 variant and the APC variant. The results from the sequencing showed that the APC variant c.3086_3087insTCGG and the STK11 variant c.459_460insAGA were both false positive. The PTEN variant c.377C>T, the POLE variants c.1373A>T and c.824A>T and the BMPR1A variant c.785T>C were confirmed to be true variants. An alignment of POLE with the position of the mutations marked can be seen below in figure 3.4.

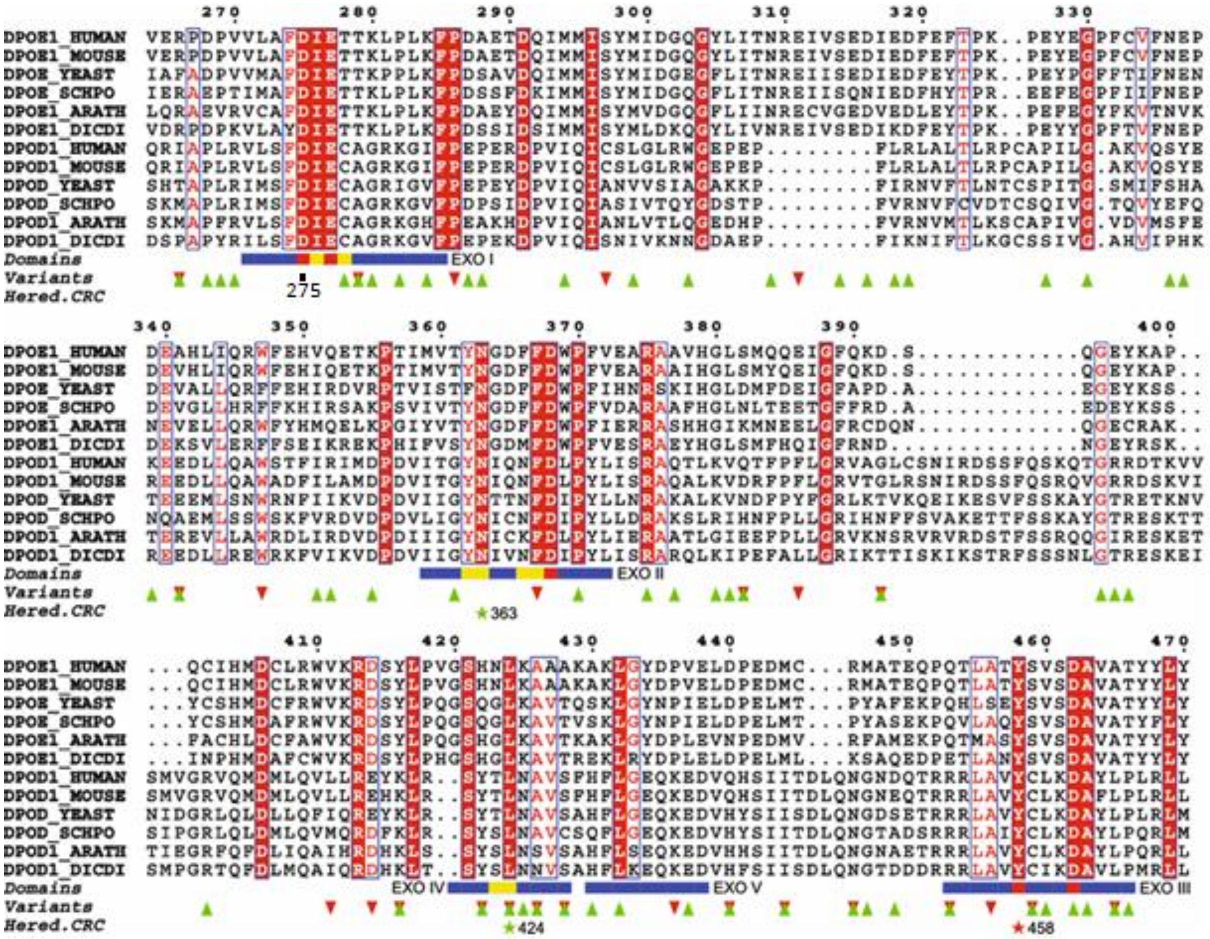


Figure 3.4. Overview of an alignment of POLE and POLD1 in several species. The two positions of the variants identified in this study p.Asp275Val and p.Tyr458Phe are marked with a black square and red star. The blue boxes show conserved positions with red background for completely conserved positions. Blue horizontal lines show the exonuclease domains and catalytic residues are shown with red squares within the exonuclease domains. [20]

4. Discussion

The purpose of this study was to detect novel variants in genes that are both well-known and not so well-known to predispose to CRC in patients that have a higher increased risk for developing CRC than the general population.

4.1 Variants found in genes known to predispose to inherited CRC syndromes

4.1.1 The variants POLE c.1373A>T and c.824A>T

The Pole variant c.1373A>T:p.Tyr458Phe that was found in three related individuals during this study was also found in an exome project of one large family by Hansen et al.[20] and it was found to be highly penetrant [20].

In the study by Hansen et al. [20] it was found that the tyrosine in this position is completely conserved between species and that the position is important for exonuclease activity. When the same position in orthologs was mutated to phenylalanine, alanine or histidine the exonuclease activity was significantly reduced, resulting in reduced fidelity of DNA replication and an increased mutation rate. The POLE variant c.1373A>T which was validated with Sanger sequencing seemed to be the cause of CRC in the family in the study by Hansen et al.[20]

The mutation in the POLE variant that causes the substitution from tyrosine to phenylalanine was in this study predicted to be damaging at protein level by all the prediction programs in Alamut. The alignment in figure 3.4 in chapter 3 shows that the mutation lies in a catalytic residue and a highly conserved region. The POLE variant identified in the samples from the patients in this study were from the same family as the one studied in Hansen et al [20].

Palles et al. [17] also found a variant in POLE and POLD1 which was heterozygous germline variants that were not found in any controls. This missense variant in POLE was a p.Leu424Val which was detected in a family with adenomas and CRC, and it appeared to have a dominant inheritance with a high penetrance as well. The change from leucine to a valine was according to the study predicted to have severe functional consequences for the protein function, including that the amino acid itself was highly conserved. By mapping the mutation found in POLE and POLD1 onto the structure of yeast DNA polymerase it was found that they pack together at the interface between two helices which form the base of the exonuclease active site. This means that mutations of POLE 424 and POLD1 478 will alter the packing of the helices and thereby distort the active site which will then affect the nuclease activity. [17]

The other variant in POLE c.824A>T:p.Asp275Val that was found in this study was identified in one patient and has not been found in any other studies. The prediction programs in Alamut predicted the mutation causing the substitution from Aspartic acid to Valine to be damaging at protein level. From the alignment in figure 3.4 it can be seen that this mutation as well lies in a catalytic residue and a highly conserved region. This suggests that both mutations identified in POLE during this study might affect the exonuclease acitivity of the protein. These findings and the findings in the studies described above strongly indicate that these variants are involved in CRC development.

4.1.2 The variant BMPR1A c.785T>C

The BMPR1A variant c.785T>C:p.Val262Ala identified in this study was found in one individual and it has not been reported earlier. The substitution from Valine to Alanine was predicted to be damaging at protein level by three of the four prediction programs in Alamut. According to this software the mutation lies in a highly conserved region indicating that alterations may affect the protein activity. On the other hand because one of the prediction programs (SIFT) in Alamut predicted the mutation to be tolerated, it is not definite that it has a damaging effect on the protein. Due to this gene’s involvement in CRC and because mutations in this gene causes JPS there is a strong possibility that the variant found in this study might also be involved in predisposition to CRC.

4.1.3 The variant PTEN c.377C>T

The PTEN c.377C>T:p.Ala126Val variant found during this study was found in two patients and have not been reported earlier. In a study by Tan et al. [58] a mutation in PTEN was found in the same codon as the variant identified in this study. The variant identified by Tan et al. [58] was a missense variant c.376G>C :p.Ala126Pro which was found to be pathogenic.

This indicates that the PTEN variant c.377C>T might also be pathogenic because it is located in the same codon. The difference between valine and proline is that proline is very unique because it is the only amino acid where the side chain is connected to the protein backbone twice. This makes proline an imino acid in its isolated form because it contains a NH^{2+} group instead of a NH^{3+} group. Due to this difference proline is unable to occupy several of the main-chain conformations which are easily adopted by the other amino acids. Proline often does not substitute well due to its unique properties.

Three of four prediction programs in Alamut predicted the variant identified in PTEN to be damaging at protein level. The mutation was found to lie in a highly conserved region which indicates that the mutation may affect protein activity. It is not definite that the protein activity will be affected due to one prediction program (SIFT) classifying the mutation as tolerated. Since germline mutations in PTEN are the cause of both CS and BRRS, it is likely that the variant identified in this project could be involved in the predisposition to CRC.

4.1.4 The variant GREM1 c.196_197insT

The variant GREM1 c.196_197insT:p.Thr66Ilefs was identified in three individuals and the variant has not been reported earlier. The Gremlin 1, DAN family BMP antagonist (GREM1; MIM # 603054) encodes a member of the bone morphogenic protein (BMP) antagonist family and they contain cystine knots and form homo-and heterodimers. The gene belongs to a subfamily of BMP anatagonists the CAN (Cerberus and dan) and is characterized by a C-terminal cystine knot with an eight-membered ring. GREM1 might be involved in regulation of organogenesis, body patterning and tissue differentiation. [59]

In a study by Jaeger et al. [60] there was identified a duplication across the 3` end of the SCG5 gene and a region upstream of the GREM1 locus in Ashkenazi Jewish families with hereditary mixed polyposis syndrome (HMPS). This syndrome has a clear autosomal dominant inheritance of several different types of colorectal polyps and the affected individuals have a high occurrence of colorectal carcinoma. The duplication contains enhancer elements where some interact with the GREM1 promoter and are able to force gene expression in vitro. The mutation identified is associated with increased allele-specific GREM1 expression and GREM1 expression can cause reduced BMP activity which is also the mechanism behind tumorigenesis in JPS. The mutation is a polymorphism rs4779584. [60]

In another study executed by Yang et al. [61] 12 case-control studies involving several cases of CRC and healthy controls the rs4779584 polymorphism was investigated to see if it was associated with CRC. The results from the study showed that the GREM1-SCG5 rs4779584 polymorphisms were associated with CRC in all the genetic models that were studied in the meta-analysis of the 12 case-control studies. The findings in the study suggest that the polymorphisms might give an increased risk for developing CRC. [61]

Due to the mutation in GREM1 being a frameshift mutation it is almost always damaging at protein level because it causes a shift in the reading frame, and this might be associated with decreased GREM1 expression. The findings in these studies indicate that the GREM1 gene might be involved in the predisposition to CRC. This means that the variant found during this study could also be involved in CRC development.

4.2 Variants associated with CRC found in GWAS and NGS

4.2.1 Variants found in FAM166A, MAML3, PPP1CB, NOTCH3, LAMB4, FANCM and RAI1

The variant identified in this study in FAM166A c.751_752del:pLeu251Valfs was found in three individuals, LAMB4 c.5265delA:p.Lys1755Asnfs was found in one individual and also found in the study by Smith et al. [62]. MAML3 c.1513_1514del:p.Gln505Alafs was identified in two individuals and MAML3 c.1506delG:p.Gln502Hisfs was found in nine patients. FANCM c.5607_5608del:p.Glu1870Aspfs was found in two patients. The variant found in RAI1 c.867_872del p.Gln290_Gln291del was identified in two patients. NOTCH3 c.3733_3734insT:p.Thr1245Ilefs and PPP1CB c.469_470insAGATC:p.Cys157 were both identified in one patient. None of the variants except the variant identified in LAMB4 have been previously reported.

In the study by Smith et al. [62] 1138 genes in 50 sporadic patients with advanced CRC were exome resequenced to find rare or novel germline mutations that were likely to play a role in colorectal tumorigenesis. The study identified germline mutations in the genes FAM166A, MAML3 (MIM # 608991), PPP1CB (MIM # 600590), NOTCH3 (MIM # 600276), LAMB4, FANCM (MIM # 609644) and RAI1 (MIM #607642). The variants in the genes FAM166A, MAML3, PPP1CB and RAI1 were not described further in the study, but according to the study they are likely to play a role in CRC. The germline mutation found in NOTCH3 by Smith et al. [62] was identified in a patient diagnosed with CRC at the age of 29. This patient had no family history of CRC. NOTCH3 has recently been found to modulate the tumorigenic properties of CRC cell, but because nonsynonymous mutations are associated with cereberal autosomal dominant arteriopathy with subcortical infacts and leukoencephalopathy (CADASIL), further studies are needed to determine if loss of protein function in this gene is associated with CRC. The mutation identified in LAMB4 was found in a patient diagnosed with CRC at the age of 68 and it showed somatic loss of the wild-type LAMB4 allele. The patient diagnosed with this variant did not have a history of other cancers but had a grandfather that died of CRC at the age of 75. Although this mutation was identified in a CRC patient it was concluded in the study that LAMB4 was not likely to play a significant role in predisposition to CRC. In FANCM both germline and somatic mutations were identified and the mutations were found in two unrelated patients with CRC. The mutations were consistent with the two-hit hypothesis and the germline mutation was also identified in one control sample. [62]

There is little information available about these genes and their role in CRC, therefore further research is necessary to determine if the variants identified in this study are involved in predisposition to CRC.

4.2.2 The variants found in BUB1B and DCC

The variants identified in BUB1B c.2252_2253insAGA:p.Pro751_Lys752insAsp and c.2253_2254insCGG:p.Pro751_Lys752insArg might be the same variant and c.2253_2254insCGG might be a false positive. The variant in BUB1B was found in seven patients. The Variant in DCC c.1664_1665insCGAGAT:p.Asn555_Gly556insGluIle was identified in one patient. These variants have not been reported earlier.

The DCC (MIM #120470) gene with the cytogenetic location 18q21.3 encodes the protein netrin 1 receptor which functions as a tumor suppressor and is often mutated or downregulated in CRC and esophageal carcinoma [63]. In a study by Popat et al. [64] it was found that patients with CRC with chromosome 18q allelic imbalance or loss of DCC expression have a poorer prognosis [64].

The BUB1B (MIM #602860) gene may cause CIN in CRC [65]. In a study by Cahill et al. [66] somatic mutations were identified in 2 out of 19 CRC cell lines and these were not identified in 40 normal alleles [66].

There is little information available about BUB1B and its function in CRC, therefore further studies are necessary to determine the function of the variant identified in this gene in CRC predisposition. DCC seems to play a role in CRC predisposition indicating that the variant identified in this gene during this study might play a role CRC. The definite function of the DCC variant in CRC is not clear, thus further research is necessary to determine this.

4.2.3 Variants found in AKT1, BUB1, KIF23, LAMC1, PIK3CA and TBX3

The variants identified in AKT1 c.206G>C:p.Arg69Pro and c.520C>T:p.Arg174Cys, KIF23 c.610_618del:p.Phe204_Lys206del, PIK3CA c.107_108insAGAT:p.Cys36fs and TBX3 c.1893del:p.Asn632Thrfs were each found in one individual. The variant identified in BUB1 c.447_448insTCT:p.Glu149_Thr150insSer was found in two patients and the variant identified in LAMC1 c. 4579_4580del:p.Leu1527Glyfs was found in four patients. Neither of the variants have been previously reported.

Mutations in these genes have in some studies been found to be associated with CRC and these will be briefly described below.

A somatic mutation in AKT1 (MIM #164730) was identified in a study by Carpten et al. [67] in human breast, ovarian and colorectal cancers. The study showed that the mutation identified activates AKT1 through pathological localization to the plasma membrane, stimulates downstream signaling, transforms cells and induces leukemia in mice. This process suggests that AKT1 has a direct role in human cancer, and adds to known genetic changes that promote oncogenesis through the phosphatidylinositol-3-OH kinase/AKT pathway. [67] In a study by Orloff et al. [68] mutations in AKT1 were found to be associated with CS. According to the results 91 probands with CS negative for mutations in the known disease-causing genes, two were found to have germline mutations in AKT1. The effect of the mutations was increased P-Thr308-AKT and increased cellular PIP3. [68]

The BUB1 (MIM # 602452) locus was in a study by Jaffrey et al. [69] studied in 32 CRC patients and in 20 non-small cell lung cancer (NSCLC) primary tumours with a panel of seven microsatellite repeats for 2q, two CA repeats in BUB1 and gene mutation analysis. In 20 of 32 colorectal primary tumors the 2q locus was quite unstable. Results also showed 14.5% of CRC patients with instability within BUB1. Jaffery et al. [69] concluded that in this study 2q and BUB1 allelic instability in CRC were shown, but mutations in BUB1 are rare causes of chromosomal instability in CRC or NSCLC. [69] In another study De Voer et al. [70] identified haploinsufficiency or heterozygous mutations in the spindle assembly checkpoint genes BUB1 and BUB3 with genome-wide and targeted copy number and mutation analysis. 208 patients with familial or early onset CRC were analyzed and they also had variegated aneuploidies in multiple tissues and variable dysmorphic features. The discoveries in this study indicated that mutations in both BUB1 and BUB3 cause mosaic variegated aneuploidy which increases the risk of CRC at a young age. [70]

A missense variant in KIF23 (MIM # 605064) was identified in the study by DeRycke et al. [7] where 40 cases from 16 familial CRC families were germline exome sequenced. It was found to be a rare variant and it was only observed in the ESP database of European Americans. The variant was validated and replicated and is located in previously reported CRC linkage regions. [7]

In a study by Peters et al. [71] a polymorphism in LAMC1 (MIM #150290) rs10911251 and in TBX3 (MIM # 601621) rs59336 were identified in GWAS. Both of the polymorphisms were associated with CRC. The polymorphism in LAMC1 is located in a region that is highly evolutionary conserved. The SNP is also close to the promoter which indicates that it might influence gene transcription. The study strongly suggested that this polymorphism is involved in development of CRC. TBX3 has also been found to be over expressed in several cancers such as in pancreatic-, liver-, breast cancer and melanoma. TBX3 was in liver cancer observed as a downstream target of the Wnt/β-catenin pathway, mediating β-catenin activities on cell proliferation and survival. This pathway is known to play an important role in CRC development. [71] The study by Whiffin et al. [72] confirmed the LAMC1 SNP's association to CRC in their meta-analysis of five GWAS. [72]

In a study by Shan et al. [73] TBX3 expression was found to be higher in CRC tissues than in normal tissues. The study suggested that TBX3 might be involved in CRC development by participating in the Epithelial-Transition Mesenchymal (EMT), and EMT have been suggested to be involved in regulation of cancer metastasis. TBX3 might also have the potential to be an effective prognostic predictor for CRC patients. [73]

Somatic mutations in PIK3CA (MIM # 171834) in 74 tumors out of 199 CRC were identified in a study by Samuels et al. [74]. The location of the mutations in PIK3CA indicated that they are likely to increase kinase activity. The results from the study suggest that PIK3CA when mutated is likely to function as an oncogene in human cancers. [74] In the study by Orloff et al. [68] 8 probands with CS who were negative for mutations in the known causing CS genes were found to have heterozygous germline mutations in the PIK3CA gene. Functional assays showed that the result of these mutations were upregulation of AKT1 phosphorylated at thr308 and increased cellular PIP3. [68].

The variants identified in AKT1 lies according to Alamut in highly conserved regions meaning that the mutations will have an effect on protein activity. The variants identified in the genes LAMC1, PIK3CA and TBX3 are frameshift mutations which means that they will also have an effect on the protein activity. The BUB1 and KIF23 variants are nonframeshift mutations and these might have an effect on the protein activity as well, although not as major as a frameshift mutation because they do not cause a shift in the reading frame. There is not sufficient enough literature available for these genes and their function in CRC to determine if the variants identified in this study are involved in predisposition to CRC. To determine the function of these genes in CRC predisposition further functional studies are needed.

4.3 Targeted NGS Sequencing

Targeted sequencing is a technique that is very useful in cancer research due to its ability to focus on specific genes and also the entire human genome. The opportunity to use genetic testing in the diagnosis of diseases with complex genetics is very valuable and in cancer research the possibility to sequence only parts of the genome and focus on specific genes is extremely helpful in the research process. [41] Targeted sequencing also enables sequencing of all CRC related genes simultaneously which increases the time it takes to assess a patient.

In this study there were detected 1268 unique variants using the Haloplex targeted NGS method where 25 of these were chosen to further investigate based on the criteria that the mutations were damaging at protein level. From the 25 variants two variants, one in APC and one in STK11, were found to be false positive by Sanger sequencing. The reason for this might be because during processing of NGS data it is not possible to remove the PCR duplicates because enzymes are used to digest the DNA. Thus removal of PCR duplicates would result in removal of parts of the PCR product. Due to only 6 variants being validated with Sanger sequencing there is a possibility that there are additional false positive variants. There were also found unequal number of variants in the samples that were from the same patients. The samples 33 and 46 that were from the same patient had a variant similarity of about 92% and the samples 51 and 87 had a variant similarity of about 95-97%. The possible explanation for the discrepancy in variants between the samples is that the variants might have been false positive and that the reason for their appearance might also be because PCR duplicates are impossible to remove. Another possible false positive variant is the second framshift insertion variant detected in BUB1B c.2253_2254insCGG because this variant seem to be in the same region as the other BUB1B variant with the only difference being one nucleotide. The 10 false positive variants detected in this study are an indication that the reproducibility of this method is not 100%. There is no guarantee that there are false negatives among the variants identified, and due to the selection criteria in this study to only focus on the frameshift variants and some missense variants known in CRC there is a possibility that highly penetrant variants have been lost.

Even though false positive variants were detected the rate of these variants are not high and because the two POLE variants, the BMPR1A and PTEN variant detected were most likely to be involved in CRC development and that several other variants detected were likely to be involved in CRC such as those found in the genes PIK3CA, AKT1, DCC, GREM1 among others, targeted sequencing seems to be a reliable method to use in cancer research.

4.4 Conclusion and prospective work

The purpose of this study was to identify novel variants in patients with an increased risk of developing CRC, using targeted NGS sequencing in genes known to be involved in hereditary CRC syndromes and in genes associated with CRC. Many variants were identified and among these two novel variants in POLE c.1373A>T and c.824A>T, the variant in BMPR1A c.785T>C and the variant in PTEN c.377C>T were found to be involved in CRC development. This proves that targeted sequencing seems to be a useful tool in identifying novel variants in CRC, but because not all variants were validated there is no guarantee that there are other false positive variants among those discovered in this study. There is also a possibility of false negative variants and loss of highly penetrant variants due to not all variants being further investigated. Prospective work will therefore be to validate the variants found in the genes AKT1, BUB1, BUB1B, DCC, FAM166A, FANCM, GREM1, KIF23, LAMB4, LAMC1, MAML3, NOTCH3, PIK3CA, PPP1CB, RAI1 and TBX3. These genes have all been associated with CRC but there is not sufficient information to state that they are involved in CRC and therefore further functional studies are needed. The variants identified in this study that were not further described mainly the missense variants, need further functional studies to determine their role in CRC development.

5. References

1. Nussbaum, R.L., et al., *Thompson & Thompson genetics in medicine*. 2007, Philadelphia: Saunders/Elsevier. XI, 585 s. : ill.

2. Patel, S.G. and D.J. Ahnen, *Familial colon cancer syndromes: an update of a rapidly evolving field.* Curr Gastroenterol Rep, 2012. **14**(5): p. 428-38.

3. Fearon, E.R., *Molecular genetics of colorectal cancer.* Annu Rev Pathol, 2011. **6**: p. 479-507.

4. Ku, C.S., et al., *Gene discovery in familial cancer syndromes by exome sequencing: prospects for the elucidation of familial colorectal cancer type X.* Mod Pathol, 2012. **25**(8): p. 1055-68.

5. Jasperson, K.W., et al., *Hereditary and familial colon cancer.* Gastroenterology, 2010. **138**(6): p. 2044-58.

6. Blanes, A. and S.J. Diaz-Cano, *Complementary analysis of microsatellite tumor profile and mismatch repair defects in colorectal carcinomas.* World J Gastroenterol, 2006. **12**(37): p. 5932-40.

7. DeRycke, M.S., et al., *Identification of novel variants in colorectal cancer families by high-throughput exome sequencing.* Cancer Epidemiol Biomarkers Prev, 2013. **22**(7): p. 1239-51.

8. Esteban-Jurado, C., et al., *New genes emerging for colorectal cancer predisposition.* World J Gastroenterol, 2014. **20**(8): p. 1961-71.

9. Vogelstein, B., et al., *Genetic alterations during colorectal-tumor development.* N Engl J Med, 1988. **319**(9): p. 525-32.

10. Oláh, E. *Basic Concepts of Cancer: Genomic Determination*. 2005 2015 [cited 2015 13/04-2015]; Available from: http://www.ifcc.org/ifcc-communications-publications-division-(cpd)/ifcc-publications/ejifcc-(journal)/e-journal-volumes/ejifcc-2005-vol-16/vol-16-n%C2%B0-2/basic-concepts-of-cancer-genomic-determination/.

11. Steinke, V., et al., *Hereditary nonpolyposis colorectal cancer (HNPCC)/Lynch syndrome.* Dtsch Arztebl Int, 2013. **110**(3): p. 32-8.

12. Knudson, A.G., *Two genetic hits (more or less) to cancer.* Nat Rev Cancer, 2001. **1**(2): p. 157-62.

13. Information, N.C.f.B. *APC adenomatous polyposis coli [Homo sapiens (human)].* 05/04-2015 [cited 2015 06/01]; Available from: http://www.ncbi.nlm.nih.gov/gene/?term=324.

14. Gala, M. and D.C. Chung, *Hereditary colon cancer syndromes.* Semin Oncol, 2011. **38**(4): p. 490-9.

15. Guarinos, C., et al., *Prevalence and characteristics of MUTYH-associated polyposis in patients with multiple adenomatous and serrated polyps.* Clin Cancer Res, 2014. **20**(5): p. 1158-68.

16. Information, N.C.f.B. *MUTYH mutY homolog [Homo sapiens (human)].* 05/04-2015 [cited 2015 10/01]; Available from: http://www.ncbi.nlm.nih.gov/gene/?term=4595.

17. Palles, C., et al., *Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas.* Nat Genet, 2013. **45**(2): p. 136-44.

18. Information, N.C.f.B. *POLE polymerase (DNA directed), epsilon, catalytic subunit [Homo sapiens (human).* 05/04-2015 [cited 2015 18/04]; Available from: http://www.ncbi.nlm.nih.gov/gene/5426.

19. Information, N.C.f.B. *POLD1 polymerase (DNA directed), delta 1, catalytic subunit [Homo sapiens (human).* 05/04-2015 [cited 2015 18/04]; Available from: http://www.ncbi.nlm.nih.gov/gene/5424.

20. Hansen, M.F., et al., *A novel POLE mutation associated with cancers of colon, pancreas, ovaries and small intestine.* Fam Cancer, 2015.

21. Manfredi, M., *Hereditary hamartomatous polyposis syndromes: understanding the disease risks as children reach adulthood.* Gastroenterol Hepatol (N Y), 2010. **6**(3): p. 185-96.

22. Beggs, A.D., et al., *Peutz-Jeghers syndrome: a systematic review and recommendations for management.* Gut, 2010. **59**(7): p. 975-86.

23. Jelsig, A.M., et al., *Hamartomatous polyposis syndromes: a review*. Orphanet J Rare Dis, 2014. **9**: p. 101.

24. Chae, H.D. and C.H. Jeon, *Peutz-Jeghers syndrome with germline mutation of STK11*. Ann Surg Treat Res, 2014. **86**(6): p. 325-30.

25. Aaltonen, L.A., *Hereditary intestinal cancer*. Semin Cancer Biol, 2000. **10**(4): p. 289-98.

26. Uppal, S., D. Mistry, and A.P. Coatesworth, *Cowden disease: a review*. Int J Clin Pract, 2007. **61**(4): p. 645-52.

27. Pezzolesi, M.G., et al., *Comparative genomic and functional analyses reveal a novel cis-acting PTEN regulatory element as a highly conserved functional E-box motif deleted in Cowden syndrome*. Hum Mol Genet, 2007. **16**(9): p. 1058-71.

28. Information, N.C.f.B. *PTEN phosphatase and tensin homolog [Homo sapiens (human)]*. 05/04-2015 [cited 2015 10/01]; Available from: http://www.ncbi.nlm.nih.gov/gene/?term=5728.

29. Gu, J., M. Tamura, and K.M. Yamada, *Tumor suppressor PTEN inhibits integrin- and growth factor-mediated mitogen-activated protein (MAP) kinase signaling pathways*. J Cell Biol, 1998. **143**(5): p. 1375-83.

30. Lammi, L., et al., *Mutations in AXIN2 cause familial tooth agenesis and predispose to colorectal cancer*. Am J Hum Genet, 2004. **74**(5): p. 1043-50.

31. Information, N.C.f.B. *AXIN2 axin 2 [Homo sapiens (human)]*. 12/04-2015 [cited 2015 17/01]; Available from: http://www.ncbi.nlm.nih.gov/gene/?term=8313.

32. Alhopuro, P., et al., *Unregulated smooth-muscle myosin in human intestinal neoplasia*. Proc Natl Acad Sci U S A, 2008. **105**(14): p. 5513-8.

33. Bennett, K.L., J. Mester, and C. Eng, *Germline epigenetic regulation of KILLIN in Cowden and Cowden-like syndrome*. Jama, 2010. **304**(24): p. 2724-31.

34. Gylfe, A.E., et al., *Eleven candidate susceptibility genes for common familial colorectal cancer*. PLoS Genet, 2013. **9**(10): p. e1003876.

35. Smith, C.G., et al., *Role of the oxidative DNA damage repair gene OGG1 in colorectal tumorigenesis.* J Natl Cancer Inst, 2013. **105**(16): p. 1249-53.

36. Kim, I.J., et al., *Mutational analysis of OGG1, MYH, MTH1 in FAP, HNPCC and sporadic colorectal cancer patients: R154H OGG1 polymorphism is associated with sporadic colorectal cancer patients.* Hum Genet, 2004. **115**(6): p. 498-503.

37. Kokko, A., et al., *EPHB2 germline variants in patients with colorectal cancer or hyperplastic polyposis.* BMC Cancer, 2006. **6**: p. 145.

38. Guda, K., et al., *Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers.* Proc Natl Acad Sci U S A, 2009. **106**(31): p. 12921-5.

39. Bamshad, M.J., et al., *Exome sequencing as a tool for Mendelian disease gene discovery.* Nat Rev Genet, 2011. **12**(11): p. 745-55.

40. Reis-Filho, J.S., *Next-generation sequencing.* Breast Cancer Res, 2009. **11 Suppl 3**: p. S12.

41. Rizzo, J.M. and M.J. Buck, *Key principles and clinical applications of "next-generation" DNA sequencing.* Cancer Prev Res (Phila), 2012. **5**(7): p. 887-900.

42. Mamanova, L., et al., *Target-enrichment strategies for next-generation sequencing.* Nat Methods, 2010. **7**(2): p. 111-8.

43. Cragun, D., et al., *Panel-based testing for inherited colorectal cancer: a descriptive study of clinical testing performed by a US laboratory.* Clin Genet, 2014. **86**(6): p. 510-20.

44. Julie, C., et al., *Identification in daily practice of patients with Lynch syndrome (hereditary nonpolyposis colorectal cancer): revised Bethesda guidelines-based approach versus molecular screening.* Am J Gastroenterol, 2008. **103**(11): p. 2825-35; quiz 2836.

45. Technologies, I.b.L. *User guide, iPrep™ PureLink® gDNA Blood Kit For purification of gDNA from human blood using the iPrep™ Purification Instrument* 2012; Revision

3.0:[Available from: http://tools.lifetechnologies.com/content/sfs/manuals/iprep_bloodgDNA_man.pdf.

46. Technologies, I.b.L. *Qubit® 2.0 Fluorometer, User Manual*. 2010 04/10-2010; Available from: https://tools.lifetechnologies.com/content/sfs/manuals/mp32866.pdf.

47. NanoDrop Technologies, I. *ND-1000 Spectrophotometer V3.2 User's Manual* 2005; Manual]. Available from: https://www.urmc.rochester.edu/fgc/documents/nd1000.pdf.

48. Technologies, A. *Haloplex Target Enrichment System For Illumina Sequencing, Protocol* 2013 May 2013; Version D.5:[Available from: http://www.chem.agilent.com/Library/usermanuals/Public/G9900-90001.pdf.

49. Technologies, A. *About Haloplex*. 2015 [cited 2015 15/01]; Available from: http://www.genomics.agilent.com/article.jsp?pageId=3267.

50. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.

51. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.* Curr Protoc Bioinformatics, 2013. **11**(1110): p. 11.10.1-11.10.33.

52. DePristo, M.A., et al., *A framework for variation discovery and genotyping using next-generation DNA sequencing data.* Nat Genet, 2011. **43**(5): p. 491-8.

53. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.

54. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.

55. Vigeland, M.D. *FILTUS*. [cited 2015 03/02]; Available from: http://folk.uio.no/magnusv/filtus.html.

56. Biosoftware, I. *Features of Alamut Visual* 2015 22/04-15 [cited 2015 06/04]; Available from: http://www.interactive-biosoftware.com/alamut-visual/features/.

57. Lyons, R. *Interpretation of Sequencing Chromatograms*. 2015 [cited 2015 08/05]; Available from: http://seqcore.brcf.med.umich.edu/doc/dnaseq/interpret.html.

58. Tan, M.H., et al., *A clinical scoring system for selection of patients for PTEN mutation testing is proposed on the basis of a prospective study of 3042 probands.* Am J Hum Genet, 2011. **88**(1): p. 42-56.

59. Information, N.C.f.B. *GREM1 gremlin 1, DAN family BMP antagonist [Homo sapiens (human)]*. 12/05-2015 [cited 2015 14/05]; Available from: http://www.ncbi.nlm.nih.gov/gene/26585.

60. Jaeger, E., et al., *Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1.* Nat Genet, 2012. **44**(6): p. 699-703.

61. Yang, H., et al., *Meta-analysis of the rs4779584 polymorphism and colorectal cancer risk.* PLoS One, 2014. **9**(2): p. e89736.

62. Smith, C.G., et al., *Exome resequencing identifies potential tumor-suppressor genes that predispose to colorectal cancer.* Hum Mutat, 2013. **34**(7): p. 1026-34.

63. Information, N.C.f.B. *DCC DCC netrin 1 receptor [Homo sapiens (human)]*. 17/05-2015 [cited 2015 25/05]; Available from: http://www.ncbi.nlm.nih.gov/gene/1630.

64. Popat, S. and R.S. Houlston, *A systematic review and meta-analysis of the relationship between chromosome 18q genotype, DCC status and colorectal cancer prognosis.* Eur J Cancer, 2005. **41**(14): p. 2060-70.

65. Ogino, S. and A. Goel, *Molecular classification and correlates in colorectal cancer.* J Mol Diagn, 2008. **10**(1): p. 13-27.

66. Cahill, D.P., et al., *Mutations of mitotic checkpoint genes in human cancers.* Nature, 1998. **392**(6673): p. 300-3.

67. Carpten, J.D., et al., *A transforming mutation in the pleckstrin homology domain of AKT1 in cancer.* Nature, 2007. **448**(7152): p. 439-44.

68. Orloff, M.S., et al., *Germline PIK3CA and AKT1 mutations in Cowden and Cowden-like syndromes.* Am J Hum Genet, 2013. **92**(1): p. 76-80.

69. Jaffrey, R.G., et al., *Genomic instability at the BUB1 locus in colorectal cancer, but not in non-small cell lung cancer.* Cancer Res, 2000. **60**(16): p. 4349-52.

70. de Voer, R.M., et al., *Germline mutations in the spindle assembly checkpoint genes BUB1 and BUB3 are risk factors for colorectal cancer.* Gastroenterology, 2013. **145**(3): p. 544-7.

71. Peters, U., et al., *Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis.* Gastroenterology, 2013. **144**(4): p. 799-807.e24.

72. Whiffin, N., et al., *Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis.* Hum Mol Genet, 2014. **23**(17): p. 4729-37.

73. Shan, Z.Z., et al., *Overexpression of Tbx3 is correlated with Epithelial-Mesenchymal Transition phenotype and predicts poor prognosis of colorectal cancer.* Am J Cancer Res, 2015. **5**(1): p. 344-53.

74. Samuels, Y., et al., *High frequency of mutations of the PIK3CA gene in human cancers.* Science, 2004. **304**(5670): p. 554.

6. Appendix

6.1 Patients fulfilling the Amsterdam criteria and/or the revised Bethesda guidelines

Patient nr.	AC	RBG
1	Positive	Positive
2	Positive	Positive
3	Positive	Positive
4	Positive	Positive
5	Negative	Positive
6	Negative	Positive
7	Positive	Positive
8	Negative	Positive
9	Positive	Positive
10	Positive	Positive
11	Positive	Positive
12	Positive	Positive
13	Positive	Positive
14	Positive	Positive
15	Positive	Positive
16	Negative	Negative
17	Positive	Positive
18	Positive	Positive
19	Positive	Positive
20	Positive	Positive
21	Positive	Positive
22	Positive	Positive
23	Positive	Positive
24	Positive	Positive
25	Positive	Positive
26	Positive	Positive
27	Positive	Positive
28	Positive	Positive
29	Positive	Positive
30	Positive	Positive
31	Positive	Positive
32	Positive	Positive
33	Positive	Positive
34	Positive	Positive
35	Positive	Positive

Patient nr.	AC	RBG
36	Positive	Positive
37	Positive	Positive
38	Positive	Positive
39	Negative	Positive
40	Negative	Positive
41	Positive	Positive
42	Positive	Positive
43	Positive	Positive
44	Positive	Positive
45	Positive	Positive
46	Positive	Positive
47	Negative	Positive
48	Positive	Positive
49	Negative	Positive
50	Negative	Negative
51	Positive	Positive
51	Positive	Positive
52	Negative	Positive
53	Positive	Positive
54	Positive	Positive
55	Positive	Positive
56	Positive	Positive
57	Positive	Positive
59	Positive	Positive
60	Positive	Positive
62	Positive	Positive
63	Positive	Positive
64	Positive	Positive
65	Positive	Positive
66	Positive	Positive
67	Positive	Positive
68	Positive	Positive
69	Negative	Negative
70	Positive	Positive
71	Positive	Positive
72	Positive	Positive
73	Positive	Positive
74	Positive	Positive
75	Negative	Positive
76	Negative	Negative
77	Positive	Positive

Patient nr.	AC	RBG
78	Positive	Positive
79	Positive	Positive
80	Positive	Positive
81	Negative	Negative
82	Negative	Positive
83	Negative	Positive
84	Negative	Positive
85	Positive	Positive
86	Positive	Positive
87	Positive	
88	Negative	
89	Negative	
90	Negative	
91	Negative	
92	Negative	
93	Negative	
94	Positive	
95	Positive	
96	Positive	

6.2 Overview of DNA concentration measurement for ND-1000 and Qubit

	Concentration Nanodrop	Concentration Qubit	Sample volume	Dilution volume
Sample nr.	**ng/µL**	**ng/µL**	**µL**	**µl**
1	21,54	28,9	7,8	37,2
2	31,59	43,4	5,2	39,8
3	13,92	27,0	8,3	36,7
4	51,48	51,0	4,4	40,6
5	30,43	32,9	6,8	38,2
6	76,41	49,3	4,6	40,4
7	28,07	37,0	6,1	38,9
8	20,10	28,4	7,9	37,1
9	8,89	8,2	27,3	17,7
10	22,46	27,2	8,3	36,7
11	24,44	22,7	9,9	35,1
12	83,29	54,0	4,2	40,8
13	46,70	50,0	4,5	40,5
14	31,67	36,2	6,2	38,8
15	22,37	19,6	11,5	33,5
16	31,16	32,3	7,0	38,0
17	36,01	50,0	4,5	40,5
18	41,41	41,8	5,4	39,6
19	52,78	45,2	5,0	40,0
20	27,03	30,8	7,3	37,7
21	31,30	28,8	7,8	37,2
22	29,10	38,7	5,8	39,2
23	29,49	26,6	8,5	36,5
24	33,99	36,2	6,2	38,8

	Concentration Nanodrop	Concentration Qubit	Sample volume	Dilution volume
Sample nr.	**ng/µL**	**ng/µL**	**µL**	**µL**
25	33,02	36,3	6,2	38,8
26	23,58	22,2	10,1	34,9
27	39,30	44,1	5,1	39,9
28	29,73	28,3	8,0	37,0
29	23,19	22,8	9,9	35,1
30	29,66	26,2	8,6	36,4
31	31,22	29,7	7,6	37,4
32	43,44	43,2	5,2	39,8
33	22,73	25,5	8,8	36,2
34	37,22	41,1	5,5	39,5
35	28,28	15,3	14,7	30,3
36	31,79	43,1	5,2	39,8
37	48,29	30,0	7,5	37,5
38	43,29	14,2	15,8	29,2
39	59,52	37,8	6,0	39,0
40	29,63	19,8	11,4	33,6
41	23,53	20,8	10,8	34,2
42	26,12	22,8	9,9	35,1
43	39,51	17,3	13,0	32,0
44	71,92	27,9	8,1	36,9
45	27,06	47,3	4,8	40,2
46	53,19	30,2	7,5	37,5
47	19,40	25,7	8,8	36,2
48	25,59	31,4	7,2	37,8
49	32,73	45,2	5,0	40,0
50	28,12	27,8	8,1	36,9

	Concentration Nanodrop	Concentration Qubit	Sample volume	Dilution volume
Sample nr.	**ng/µL**	**ng/µL**	**µL**	**µL**
51	98,33	23,1	9,7	35,3
52	52,57	27,6	8,2	36,8
53	32,06	29,2	7,7	37,3
54	39,64	21,8	10,3	34,7
55	54,57	43,3	5,2	39,8
56	46,12	46,2	4,9	40,1
57	22,72	11,6	19,4	25,6
58	**CONTROL**	**CONTROL**	**CONTROL**	**CONTROL**
59	74,04	48,6	4,6	40,4
60	31,16	35,0	6,4	38,6
61	55,81	56,0	4,0	41,0
62	36,77	51,0	4,4	40,6
63	28,35	32,3	7,0	38,0
64	55,30	38,7	5,8	39,2
65	31,53	38,7	5,8	39,2
66	49,70	35,5	6,3	38,7
67	48,63	46,1	4,9	40,1
68	38,00	36,4	6,2	38,8
69	62,23	42,3	5,3	39,7
70	43,86	39,7	5,7	39,3
71	51,78	40,3	5,6	39,4
72	21,75	22,6	10,0	35,0
73	13,54	17,1	13,2	31,8
74	24,29	33,1	6,8	38,2
75	83,05	55,0	4,1	40,9
76	55,39	51,0	4,4	40,6

	Concentration Nanodrop	Concentration Qubit	Sample volume	Dilution volume
Sample nr.	ng/µL	ng/µL	µL	µL
77	67,96	56,0	4,0	41,0
78	64,76	59,0	3,8	41,2
79	27,42	21,1	10,7	34,3
80	58,20	54,0	4,2	40,8
81	110,08	58,0	3,9	41,1
82	74,62	53,0	4,2	40,8
83	75,64	46,4	4,8	40,2
84	42,41	40,0	5,6	39,4
85	110,38	112,0	2,0	43,0
86	32,51	29,4	7,7	37,3
87	56,91	51,0	4,4	40,6
88	99,49	40,9	5,5	39,5
89	35,64	32,0	7,0	38,0
90	74,65	42,8	5,3	39,7
91	57,37	40,2	5,6	39,4
92	84,39	65,4	3,4	41,6
93	62,71	49,5	4,5	40,5
94	40,19	39,7	5,7	39,3
95	46,75	32,6	6,9	38,1
96	56,14	27,2	8,3	36,7

6.3 Bioanalyzer results before and after pooling of samples

The table presents results from validation of enrichment and quantity of enriched target DNA for each sample measured on 2100 Bioanalyzer. The table also shows the volume of each sample to obtain equimolar concentration prior to pooling. The expected- and the actual concentration in the pool can also be seen.

	Sample nr	**Concentration (ng/µl)**	**Volume for equimolar concentration**
Chip 1	1	4,53	2,21
	2	5,56	1,80
	3	3,18	3,14
	4	4,28	2,34
	5	4,45	2,25
	6	9,42	1,06
	7	4,55	2,20
	8	1,08	9,26
	9	5,06	1,98
	10	2,55	3,92
	11	3,70	2,70
Chip 2	12	7,66	1,31
	13	3,40	2,94
	14	3,42	2,92
	15	2,58	3,88
	16	3,34	2,99
	17	5,32	1,88
	18	5,93	1,69
	19	3,17	3,15
	20	4,03	2,48
	21	2,24	4,46
	22	4,11	2,43

	Sample nr	Concentration (ng/µl)	Volume for equimolar concentration
Chip 3	23	5,55	1,80
	24	3,89	2,57
	25	3,17	3,15
	26	2,54	3,94
	27	2,23	4,48
	28	3,24	3,09
	29	3,16	3,16
	30	3,62	2,76
	31	5,19	1,93
	32	5,76	1,74
	33	1,61	6,21
Chip 4	34	4,31	2,32
	35	4,30	2,33
	36	2,44	4,10
	37	4,67	2,14
	38	5,16	1,94
	39	2,84	3,52
	40	3,90	2,56
	41	3,31	3,02
	42	4,22	2,37
	43	7,30	1,37
	44	4,26	2,35
Chip 5	45	2,76	3,62
	46	3,94	2,54
	47	4,09	2,44
	48	4,99	2,00

	Sample nr	Concentration (ng/µl)	Volume for equimolar concentration
	49	3,29	3,04
	50	4,09	2,44
	51	4,72	2,12
	52	4,84	2,07
	53	4,44	2,25
	54	3,53	2,83
	55	5,02	1,99
Chip 6	56	6,40	1,56
	57	4,01	2,49
CONTROL	58	3,69	2,71
	59	6,63	1,51
	60	3,17	3,15
	61	3,95	2,53
	62	2,74	3,65
	63	3,26	3,07
	64	5,27	1,90
	65	4,68	2,14
	66	5,94	1,68
Chip 7	67	1,30	7,69
	68	3,99	2,51
	69	4,60	2,17
	70	5,02	1,99
	71	4,22	2,37
	72	2,98	3,36
	73	2,4	4,17
	74	3,23	3,10

	Sample nr	Concentration (ng/µl)	Volume for equimolar concentration
	75	3,64	2,75
	76	5,96	1,68
	77	3,28	3,05
Chip 8	78	3,97	2,52
	79	3,67	2,72
	80	4,40	2,27
	81	6,86	1,46
	82	3,43	2,92
	83	5,45	1,83
	84	3,48	2,87
	85	2,77	3,61
	86	2,30	4,35
	87	3,55	2,82
	88	3,78	2,65
Chip 9	89	4,02	2,49
	90	4,29	2,33
	91	6,81	1,47
	92	2,78	3,60
	93	3,85	2,60
	94	4,26	2,35
	95	3,55	2,82
	96	3,01	3,32
	Sum		**265,41**
	ng in Pool	**Expected concentration of pool ng/µl**	
	950	3,58	

	Measurements **Pooled samples**	**Actual concentration of pool ng/µl**	
	1	3,93	
	2	3,62	
	3	4,01	

6.4 Overview of the 123 genes, the MMR genes are highlighted in red

Gene	References	Association with syndrome	Comment
ACVRL1			
AKR1C4	(Gylfe et al., 2013)[34]	FCC	
AKT1	MIM164730	CS	
APC	MIM611731		Velkjent predisposisjons gen
ATM			
AURKA	MIM603072		
AXIN1		Finner ingen bevis	DeRyke[7] mener kjent gen
AXIN2	MIM604025	OCCS	
BAX	MIM600040		
BCLAF1			
BGLAP			
BLM	MIM604610	BLM	Haploinsufficiency
BMP2			
BMP4			
BMPR1A	MIM601299		Velkjent predisposisjons gen
BRCA1	MIM113705		Velkjent predisposisjons gen
BRCA2	MIM600185		Velkjent predisposisjons gen
BUB1	(de Voer et al., 2013)[70]		
BUB1B	MIM602860		
BUB3	(de Voer et al., 2013)		
CCDC18	(Gylfe et al., 2013)	FCC	
CCND1	MIM168461		
CCND2			
CDH1	MIM192090	CRC +HDGC	Velkjent predisposisjons gen
CDKN1A			

Gene	References	Association with syndrome	Comment
CENPE	(DeRycke et al.,2013)	FCC	
CHEK2	MIM604373		
CTNNB1			
DCC	MIM120470		
DCLRE1A			
DSG4			
DUSP10			
DUSP4			
EIF3C			
EIF3H			
ENG	MIM131195	JPS	
EPCAM	MIM185535		
EPHB2	(Kokko et al., 2006)[37]	FCC	
EXO1	MIM606063		
FAM166A			
FANCD2			
FANCM			
FLCN	MIM607273	BHDS	
GALNT12	MIM608812	CRC	
GREM1	MIM603054		
HELQ			
KIF23	(DeRycke et al.,2013)	FCC	
KIT	MIM164920	FGST	
KLLN	MIM612105	CS	Germline epigenetic regulation (methylation)
LAMA3			
LAMA5			

Gene	References	Association with syndrome	Comment
LAMB4			
LAMC1			
LAMC3			
LIG1			
LUC7L			
MAML3			
MCC	MIM159350	Finner ingen bevis	DeRyke mener kjent gen
MLH1	MIM120436	LS	Velkjent predisposisjons gen
MLH3	MIM604395	LS	
MRPL3	(Gylfe et al., 2013)	FCC	
MSH2	MIM609309	LS	Velkjent predisposisjons gen
MSH3	(Duraturo et al.,2011)	LS	Low-risk allele
MSH6	MIM600678	LS	Velkjent predisposisjons gen
MUTYH	MIM604933		Velkjent predisposisjons gen
MYC			
MYH11	MIM160745	PJS	Recessive inheritance
NABP1			
NOTCH3			
NUDT7	(Gylfe et al., 2013)	FCC	
OGG1	(Smith et al., 2013)(Kim et al., 2004)[35, 36]	FCC	Low-risk allele
PCNA			
PICALM			
PIK3CA	MIM171834	CS	
PITX1			
PLA2G2A	MIM172411		
PMS1	Ingen bevis i OMIM		DeRyke mener kjent gen

Gene	References	Association with syndrome	Comment
PMS2	MIM600259	LS	Velkjent predisposisjons gen
PMS2CL			
POLD1	(Esteban-Jurado et al., 2014)(Palles et al., 2013)		
POLD2			
POLD3			
POLD4			
POLE	(Esteban-Jurado et al., 2014)(Palles et al., 2013)[8, 17]		
PPP1CB			
PRADC1	(Gylfe et al., 2013)	FCC	
PRSS37	(Gylfe et al., 2013)	FCC	
PSPH	(Gylfe et al., 2013)	FCC	
PTCHD3			
PTEN	MIM601728	CS	
PTPRJ	MIM600925		
RAI1			
RFC1			
RFC2			
RFC3			
RFC4			
RFC5			
RHPN2			
RPA1			
RPA2			
RPA3			

Gene	References	Association with syndrome	Comment
SFXN4	(Gylfe et al., 2013)	FCC	
SHROOM2			
SLC5A9			
SMAD4	MIM600993		Velkjent predisposisjons gen
SMAD7	MIM602932		
STK11	MIM602216		
TBX3			
TERC			
TERT			
TGFBR2	MIM190182		
TKT			
TLR2	MIM603028		
TLR4	MIM603030		
TP53	MIM191170		
TRA2A			
TREX2			
TWSG1	(Gylfe et al., 2013)	FCC	
UACA	(Gylfe et al., 2013)	FCC	
UBAP2			
USP6NL			
ZFP14			
ZMYM5			
ZNF490	(Gylfe et al., 2013)	FCC	

6.5 64 variants colour coded according to predictions with Alamut

A list of variants after filtration with information on each variant from Alamut.
The variants in red have a mutation that is predicted of all programs to be damaging to the protein, the variants colour coded with yellow have a mutation that can be damaging to the protein but the prediction programs in Alamut are contradictory and the variants in green have a mutation that is not damaging for the protein. The prediction programs in Alamut were not able to collect information about the frameshift mutations and that is the reason why N/A (not applicable) is written in these fields.

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
FANCD2	30	NM_033084.3 :c.1279G>T:p. V427F	Class C0	Deleterious (score 0.01)	Disease causing (P-value:1)	Probably damaging(Humdiv 1.00 +humvar pred 0,997)
KIF23	71	NM_138555.2 c.610_618del p.Phe204_Lys206del	N/A	N/A	N/A	N/A
KIF23	6	NM_138555.2 c.622G>C p.Glu208Gln	Class C0	Tolearated (score 0.26)	Disease causing (P-value: 1)	Possibly damaging (Humdiv 0.873 + humvar 0.588)
LAMA3	80	NM_198129.1 c.8693A>G	Class C45	Deleterious (score 0)	Disease causing (P-value: 1)	Probably damaging(Humdiv 0.999 +humvar pred 0.996)
RAI1	35,48	NM_030665.3 c.867_872del p.Gln290_Gln291del	N/A	N/A	N/A	N/A
LAMA5	3	NM_005560:c. 9691C>T:p.P3231S	Class C0	Tolerated (score 0.9)	Polymorphism (P-value: 1)	Benign (Humdiv 0.002 + humvar 0.003)
LAMA5	13	NM_005560:c. 7655C>T:p.T2552M	Class C0	Tolerated (score 0.23)	Polymorphism (P-value: 1)	Benign (Humdiv 0.029 + humvar 0.002)

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
LAMA5	29	NM_005560:c.8822C>T:p.T2941M	Class C0	Tolerated (score 0.05)	Disease causing (P-value: 0.997)	Probably damaging (Humdiv 1.000 + humvar 0.962)
LAMA5	38	NM_005560:c.2918C>T:p.T973M	Class C0	Deleterious (score 0.03)	Disease causing (P-value: 0.95)	Possibly damaging(Humdiv 0.830) Benign(humvar 0.177)
LAMA5	52	NM_005560:c.3575T>C:p.I1192T	Class C0	Tolerated (score 0.31)	Polymorphism (P-value: 1)	Benign(Humdiv 0.023 + humvar 0.010)
LAMA5	80	NM_005560:c.11015G>A:p.R3672Q	Class C0	Tolerated (score 0.41)	Polymorphism (P-value: 1)	Benign(Humdiv 0.004 + humvar 0.008)
LAMC1	14,51,71,79	NM_002293:c.4579_4580del:p.Leu1527Glyfs*7	N/A	N/A	N/A	N/A
LAMC1	77	NM_002293:c.2426A>G:p.D809G	Class C65	Deleterious	Disease causing (P-value: 1)	Probably damaging (Humdiv 1.000 + humvar 1.000)
BUB1B	11,45,50,62	NM_001211:c.2252_2253insAGA:p.Pro751_Lys752insAsp	N/A	N/A	N/A	N/A
BUB1B	11,45,62	NM_001211:c.2253_2254insCGG:p.Pro751_Lys752insArg	N/A	N/A	N/A	N/A
BUB1B	40	NM_001211:c.800A>C:p.Q267P	Class C0	Tolerated	Polymorphism (P-value: 1)	Benign (Humdiv 0.000 + humvar 0.000)
MAML3	3,37	NM_018717:c.1513_1514del:p.Gln505Alafs*21	N/A	N/A	N/A	N/A

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
MAML3	3,18,34,41,52,59,64,77,90	NM_018717:c.1506delG:p.Gln502Hisfs*20	N/A	N/A	N/A	N/A
MAML3	31	NM_018717:c.755T>G:p.I252S	Class C0	Deleterious (score 0)	Disease causing (P-value: 0.874)	Benign (Humdiv 0.090 + humvar 0.046)
MAML3	59	NM_018717:c.1713G>C:p.M571I				
MAML3	92	NM_018717:c.53T>C:p.I18T	Class C0	Deleterious (Score 0)	Disease causing (P-value: 0.991)	Possibly damaging (Humdiv 0.952 + humvar 0.521)
APC	14,24,46,47,49,59-61,72,95	NM_001127511:c.3086_3087insTCGG:p.Lys1030Argfs*2	N/A	N/A	N/A	N/A
POLE	5,23,39	NM_006231:c.229C>T:p.R77C	Class C0	Deleterious (score 0.02)	Disease causing (P-value: 1)	Probably damaging (Humdiv 0.997) Possibly damaging (humvar0.696)
POLE	6,29,30	NM_006231:c.1373A>T:p.Y458F	Class C0	Deleterious (score 0)	Disease causing (P-value: 1)	Probably damaging (Humdiv 1.000 + humvar 1.000)
POLE	44	NM_006231:c.824A>T:p.D275V	Class C0	Deleterious (score 0)	Disease causing (P-value:1)	Probably damaging (Humdiv 1.000 + humvar 1.000)
POLE	51	NM_006231:c.2644A>G:p.N882D	Class C0	Tolerated (score 0.07)	Disease causing (P-value: 0.997)	Benign (Humdiv 0.001 + humvar 0.007)

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
POLE	95	NM_006231:c.4307G>A:p.R1436Q	Class C0	Deleterious (score 0.03)	Disease causing (P-value: 1)	Probably damaging (Humdiv 0.970) Possibly damaging (humvar 0.522)
GREM1	30,31,60	NM_013372:c.196_197insT:p.Thr66Ilefs*35	N/A	N/A	N/A	N/A
LAMB4	78	NM_007356:c.5265delA:p.Lys1755Asnfs*11	N/A	N/A	N/A	N/A
UBAP2	41	NM_018449:c.212G>T:p.C71F	Class C0	Deleterious (score 0.01)	Disease causing (P-value: 1)	Probably damaging (Humdiv 1.000 + humvar 0.995)
UBAP2	55	NM_001282530:c.218G>A:p.R73Q				
UBAP2	65	NM_001282529:c.596G>A:p.R199Q				
PTCHD3	44	NM_001034842:c.1853A>G:p.Y618C	Class C0	Deleterious (score 0.02)	Disease causing (P-value: 0.986)	Probably damaging (Humdiv 0.996 + humvar 0.997)
AXIN2	27	NM_004655:c.769G>T:p.A257S	Class C0	Tolerated (score 0.07)	Disease causing (P-value: 0.999)	Benign (Humdiv 0.212 + humvar 0.040)
FAM166A	16,22,52	NM_001001710:c.751_752del:p.Leu251Valfs*2	N/A	N/A	N/A	N/A
NOTCH3	34	NM_000435:c.3733_3734insT:p.Thr1245Ilefs*20	N/A	N/A	N/A	N/A

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
NOTCH3	40	NM_000435:c.6532C>T:p.P2178S	Class C0	Tolerated (score 0.53)	Disease causing (P-value: 0.64)	Benign (Humdiv 0.012 + humvar 0.006)
NOTCH3	86	NM_000435:c.2953C>T:p.R985C	Class C35	Deleterious (score 0.01)	Disease causing (P-value: 0.999)	Probably damaging (Humdiv 0.994) Possibly damaging (humvar 0.726)
ATM	65	NM_000051:c.7308A>C:p.R2436S	Class C65	Deleterious (score 0)	Disease causing (P-value: 1)	Probably damaging (Humdiv 0.998 + humvar 0.966)
DCC	23	NM_005215:c.1817C>G:p.P606R	Class C0	Deleterious (score 0)	N/A	Probably damaging (Humdiv 0.999 + humvar 0.997)
DCC	46	NM_005215:c.1664_1665insCGAGAT:p.Asn555_Gly556insGluIle	N/A	N/A	N/A	N/A
AKT1	32	NM_001014431:c.206G>C:p.R69P	Class C35	Deleterious (score 0.01)	Disease causing (P-value: 0.995)	Possibly damaging (Humdiv 0.792 + humvar 0.667)
AKT1	46	NM_001014431:c.520C>T:p.R174C	Class C0	Deleterious (score 0)	Disease causing (P-Value: 1)	Possibly damaging (Humdiv 0.900 + humvar 0.800)
FANCM	17,28	NM_020937:c.5607_5608del:p.Glu1870Aspfs*4	N/A	N/A	N/A	N/A
BUB1	49,65	NM_ 004336.4: c. 447_448insTCT	N/A	N/A	N/A	N/A

		p.Glu149_Thr150insSer				
Gene	**Patients**	**Nomenclature**	**AlignGVGD**	**SIFT**	**MutationTaster**	**Polyphen-2**
PTEN	35,48	NM_000314:c.377C>T:p.A126V	Class C0	Tolerated (score 0.29)	Disease causing (P-value: 1)	Probably damaging (Humdiv 1.000 + humvar 0.998)
BRCA1	49	NM_ 007300.3. c.889A>G p.Met29Val	Class C0	Deleterious (score 0.04)	Polymorphism (P-value: 0.946)	Benign (Humdiv 0.074 + humvar 0.041)
STK11	60	NM_000455:c.459_460insAGA:p.Ala153_His154insArg	N/A	N/A	N/A	N/A
ZFP14	14	NM_020917:c.43T>C:p.F15L	Class C0	Deleterious (score 0)	Disease causing (P-value: 0.75)	Probably damaging (Humdiv 0.999 + humvar 0.914)
SLC5A9	66	NM_001135181.1. c.1194G>T p.Leu398Phe	Class C15	Deleterious (score 0)	Disease causing (P-value: 1)	Probably damaging (Humdiv 1.000 + humvar 0.999)
OGG1	26	NM_016821.2:c.412A>G:p.I138V	Class C0	Tolerated (score 1)	Polymorphism (P-value: 1)	Benign (Humdiv 0.000 + humvar 0.002)
PIK3CA	38	NM_006218:c.107_108insAGAT:p.Cys36fs*	N/A	N/A	N/A	N/A
TBX3	66	NM_016569.3. c.1893del p.Asn632Thrfs*257	N/A	N/A	N/A	N/A
DUSP10	85	NM_007207:c.868C>A:p.L290I	Class C0	Toleararted (score 0.35)	Disease causing (P-value: 1)	Possibly damaging (Humdiv 0.907) Benign (Humvar 0.346)

Gene	Patients	Nomenclature	AlignGVGD	SIFT	MutationTaster	Polyphen-2
BMPR1A	32	NM_004329:c.785T>C:p.V262A	Class C0	Tolerated (score 0.15)	Disease causing (P-value: 1)	Possibly damaging (Humdiv 0.923 + humvar 0.884)
TGFBR2	10	NM_001024847.2. c.1292C>T p.Pro431Leu	Class C15	Deleterious (score 0.01)	Disease causing (P-value: 1)	Benign (Humdiv 0.117 + humvar 0.123)
FLCN	74	NM_144997:c.1508G>C:p.C503S	Class C0	Tolerated (score 0.39)	Disease causing (P-value: 1)	Possibly damaging (Humdiv 0.907) Benign (humvar 0.201)
BGLAP	80	NM_199173:c.217G>A:p.V73M	Class C0	Tolerated (score 0.24)	Polymorphism (P-value: 0.739)	Possibly damaging (Humdiv 0.779) Benign (humvar 0.304)
LAMC3	61	NM_006059:c.1145C>T:p.P382L	Class C0	Tolerated (score 0.05)	Disease causing (P-value: 1)	Probably damaging (Humdiv 0.999 + humvar 0.980)
USP6NL	81	NM_001080491:c.1562T>C:p.M521T	Class C0	Tolerated (score 0.14)	Polymorphism (P-value: 1)	Benign (Humdiv 0.000 + humvar 0.000)
PPP1CB	72	NM_002709:c.469_470insAGATC:p.Cys157*	N/A	N/A	N/A	N/A
KLLN	58	NM_001126049:c.454C>T:p.P152S	Class C65	Deleterious (score 0)	Polymorphism (P-value: 1)	Possibly Damaging (Humdiv 0.728) Benign (humvar 0.358)
TLR2	29	NM_003264:c.728C>A:p.S243Y	Class C0	Tolerated (score 0.21)	Polymorphism (P-value: 1)	Benign (Humdiv 0.240 + humvar 0.061)

Gene	**Patients**	**Nomenclature**	**AlignGVGD**	**SIFT**	**MutationTaster**	**Polyphen-2**
PTPRJ	67	NM_002843:c.2017G>T:p.V673L	Class C0	Tolerated (score 0.76)	Polymorphism (P-value: 1)	Benign (Humdiv 0.000 + humvar 0.004)

6.6 List of mean coverage for each sample

An overview of mean coverage for each sample and mean coverage of total samples

Sample nr.	Coverage mean
1	217,91
2	155,14
3	284,05
4	316,01
5	225,70
6	223,10
7	248,41
8	239,04
9	219,28
10	377,72
11	262,86
12	238,36
13	272,21
14	302,66
15	294,23
16	254,96
17	213,41
18	201,69
19	279,53
20	281,70
21	322,03
22	270,34
23	245,22
24	270,31

Sample nr.	Coverage mean
25	203,44
26	335,31
27	267,49
28	244,50
29	295,02
30	291,44
31	226,33
32	118,72
33	452,88
34	248,16
35	260,02
36	233,52
37	229,98
38	482,46
39	260,85
40	301,82
41	272,62
42	312,87
43	168,09
44	272,11
45	241,66
46	259,53
47	226,98
48	157,87
49	225,70
50	210,06

Sample nr.	Coverage mean
51	183,92
52	240,90
53	185,95
54	222,56
55	238,25
56	206,00
57	207,38
59	278,21
60	250,28
61	283,00
62	332,20
63	247,79
64	199,13
65	194,97
66	230,95
67	429,97
68	238,43
69	258,52
70	214,06
71	242,11
72	307,74
73	289,34
74	280,99
75	362,32
76	167,66
77	211,47

Sample nr.	Coverage mean
78	235,08
79	234,21
80	191,96
81	285,06
82	316,37
83	237,52
84	274,00
85	272,22
86	257,62
87	317,13
88	291,38
89	217,64
90	243,84
91	225,68
92	243,00
93	302,79
94	234,70
95	298,84
96	329,03
Sum:	24527,47
Mean coverage	258,18
Standard deviation	57,76

6.7 Target regions covered with >20 reads

Overview of % target regions covered with >20 reads

Sample	#regions	#regions_covered>20	Percentage
1	2493	2137	85,72
2	2493	2065	82,83
3	2493	2141	85,88
4	2493	2168	86,96
5	2493	2114	84,80
6	2493	2173	87,16
7	2493	2140	85,84
8	2493	2157	86,52
9	2493	2168	86,96
10	2493	2209	88,61
11	2493	2136	85,68
12	2493	2158	86,56
13	2493	2161	86,68
14	2493	2172	87,12
15	2493	2129	85,40
16	2493	2170	87,04
17	2493	2152	86,32
18	2493	2129	85,40
19	2493	2179	87,40
20	2493	2198	88,17
21	2493	2204	88,41
22	2493	2212	88,73
23	2493	2175	87,24
24	2493	2207	88,53
25	2493	2094	84,00
26	2493	2172	87,12

Sample	#regions	#regions_covered>20	Percentage
27	2493	2106	84,48
28	2493	2147	86,12
29	2493	2146	86,08
30	2493	2162	86,72
31	2493	2109	84,60
32	2493	2008	80,55
33	2493	2232	89,53
34	2493	2157	86,52
35	2493	2161	86,68
36	2493	2138	85,76
37	2493	2154	86,40
38	2493	2278	91,38
39	2493	2188	87,77
40	2493	2216	88,89
41	2493	2200	88,25
42	2493	2239	89,81
43	2493	2117	84,92
44	2493	2207	88,53
45	2493	2146	86,08
46	2493	2177	87,32
47	2493	2179	87,40
48	2493	2084	83,59
49	2493	2150	86,24
50	2493	2139	85,80
51	2493	2075	83,23
52	2493	2143	85,96
53	2493	2061	82,67
54	2493	2124	85,20

Sample	#regions	#regions_covered>20	Percentage
55	2493	2158	86,56
56	2493	2139	85,80
57	2493	2126	85,28
59	2493	2184	87,61
60	2493	2140	85,84
61	2493	2167	86,92
62	2493	2235	89,65
63	2493	2183	87,57
64	2493	2144	86,00
65	2493	2142	85,92
66	2493	2178	87,36
67	2493	2265	90,85
68	2493	2170	87,04
69	2493	2210	88,65
70	2493	2161	86,68
71	2493	2141	85,88
72	2493	2198	88,17
73	2493	2183	87,57
74	2493	2161	86,68
75	2493	2204	88,41
76	2493	2087	83,71
77	2493	2132	85,52
78	2493	2150	86,24
79	2493	2138	85,76
80	2493	2103	84,36
81	2493	2174	87,20
82	2493	2210	88,65
83	2493	2145	86,04

Sample	**#regions**	**#regions_covered>20**	**Percentage**
84	2493	2173	87,16
85	2493	2184	87,61
86	2493	2178	87,36
87	2493	2204	88,41
88	2493	2194	88,01
89	2493	2171	87,08
90	2493	2166	86,88
91	2493	2141	85,88
92	2493	2159	86,60
93	2493	2210	88,65
94	2493	2178	87,36
95	2493	2227	89,33
96	2493	2196	88,09
Sum:		205372	8237,95
Mean:		2162	86,72
Standard deviation:		42,92	

6.8 Number of variants found in each patient before filtration

Patient nr.	Number of Variants	In total genes
1	295	78
2	345	83
3	305	85
4	319	80
5	297	78
6	334	83
7	319	85
8	314	84
9	338	83
10	272	76
11	321	84
12	327	74
13	310	81
14	316	83
15	321	80
16	329	81
17	319	85
18	341	81
19	344	79
20	332	88
21	330	84
22	317	88
23	312	86
24	317	83
25	326	84

Patient nr.	Number of Variants	In total genes
26	318	79
27	296	74
28	323	82
29	306	78
30	304	85
31	311	78
32	351	85
33	305	80
34	303	78
35	315	87
36	344	81
37	340	85
38	285	77
39	306	79
40	302	78
41	301	78
42	293	79
43	338	83
44	329	83
45	316	83
46	338	77
47	325	84
48	332	81
49	318	76
50	324	82
51	316	78
52	315	86

Patient nr.	Number of Variants	In total genes
53	304	77
54	335	78
55	333	81
56	328	81
57	357	85
59	319	87
60	326	82
61	298	80
62	304	82
63	320	81
64	318	82
65	320	80
66	318	85
67	272	75
68	332	79
69	319	79
70	335	80
71	321	81
72	301	82
73	317	86
74	306	79
75	309	85
76	340	83
77	331	82
78	348	80
79	316	81
80	337	80

Patient nr.	Number of Variants	In total genes
81	324	80
82	306	75
83	340	90
84	324	82
85	336	78
86	298	72
87	297	78
88	296	78
89	304	77
90	300	77
91	323	81
92	297	84
93	305	83
94	347	86
95	309	82
96	284	79

6.9 Number of variants found in each patient after filtration

Patient nr.	Number of variants	In total genes
1	17	12
2	16	13
3	15	13
4	14	13
5	14	12
6	18	14
7	16	14
8	16	13
9	21	15
10	13	10
11	16	12
12	19	16
13	15	13
14	16	13
15	14	12
16	18	15
17	19	15
18	17	14
19	21	16
20	18	15
21	14	12
22	21	16
23	17	14
24	16	13
25	19	15
26	17	14
27	17	14
28	20	16
29	18	15
30	15	13
31	19	15
32	18	14
33	14	11
34	15	13
35	18	15
36	19	14
37	18	15

Patient nr.	Number of variants	In total genes
38	14	12
39	20	16
40	17	14
41	15	12
42	21	17
43	21	15
44	21	16
45	18	15
46	18	14
47	24	19
48	17	12
49	17	13
50	24	18
51	21	17
52	17	14
53	14	12
54	17	14
55	20	15
56	22	16
57	21	16
59	24	19
60	17	14
61	17	13
62	11	9
63	17	14
64	21	16
65	20	16
66	21	16
67	11	9
68	18	15
69	16	12
70	15	12
71	21	17
72	12	10
73	18	14
74	15	12
75	13	11
76	20	15
77	21	16
78	18	13

Patient nr.	Number of variants	In total genes
79	21	18
80	16	13
81	16	13
82	16	13
83	16	13
84	15	12
85	17	14
86	15	12
87	17	14
88	16	13
89	21	16
90	18	14
91	17	14
92	19	15
93	17	13
94	23	18
95	19	14
96	13	9