Hereditary Colorectal Cancer

Identification, Characterization and Classification of Mutations

Anna Rohlin

Department of Medical and Clinical Genetics
Institute of Biomedicine
Sahlgrenska Academy at University of Gothenburg

UNIVERSITY OF GOTHENBURG

Cover illustration: Emma Nordin

Hereditary Colorectal Cancer

ISBN 978-91-628-9210-4.
(e-pub) ISBN 978-91-628-9213-5
e-published:http://hdl.handle.net/2077/37108

anna.rohlin@gu.se

Department of Medical and Clinical Genetics
Institute of Biomedicine
The Sahlgrenska Academy at University of Gothenburg

Printed by Ineko,Gothenburg, Sweden 2014

To my family

The way to find the needle in the haystack is to sit down
(Beryl Markham)

ABSTRACT

Hereditary Colorectal Cancer; Identification, Characterization and Classification of Mutations

Anna Rohlin

Hereditary factors are thought to play are role in 20-30% of all colorectal cancers. Around 6% are found as high penetrant disease-causing mutations in genes correlated to hereditary polyposis or hereditary non-polyposis syndromes. The aim was to identify new causative genes and variants and also new mutation mechanisms in families presenting with a polyposis, atypical polyposis or non-polyposis CRC phenotype.

In classical familial adenomatous polyposis (FAP) 100% of the disease-causing mutations were found in patients from the Swedish Polyposis Registry. The mutation underlying the lowered expression of the *APC* gene in one family was identified by SNP array analysis, the mutation was a split deletion of 61Kb including half of the promoter 1B. Investigation of the significance of this promoter for expression of the *APC* gene demonstrated considerable higher expression compared with the well-known promoter 1A. In order to establish a sensitive method for mosaic-mutation detection a comparison of mutation detection methods was performed. Low-frequency mosaic mutations were detected down to 1% by use of massively parallel sequencing (MPS).

Whole exome sequencing in four families with attenuated FAP (AFAP), atypical polyposis or non-polyposis syndromes identified two high penetrant disease-causing mutations. One was found in the upstream regulatory region of *GREM1* and the other in the exonuclease domain of *POLE*. Variants in low-penetrant genes possibly contributing to CRC development were also proposed from the exome sequencing and gene specific analyses of 107 patients. Sixty-seven of these patients were analyzed in a panel of 19 selected CRC predisposing genes. Truncating mutations were found in the *BMPR1A* and *SMAD4* genes in patients with a classical FAP, atypical FAP or non-polyposis phenotype. Classification of non-synonymous variants found was also performed.

In summary, using a combination of different molecular screening techniques 100 % of disease-causing mutations in classical FAP can be found. With MPS it is possible to detect low-frequency mosaic mutations down to 1 % by absolute quantification. Whole exome analyses identified mutations in the new causative genes *POLE* and *GREM1*. It was also concluded that patients without identified mutations based on phenotypical CRC classification can have mutations in genes not included in the primary routine analysis. These results will lead to improved mutation detection analysis for diagnostics and carrier testing.

Keywords: Hereditary colorectal cancer, FAP, AFAP, atypical polyposis, PPAP, mutation, APC, POLE, GREM1, exome sequencing, massively parallel sequencing, mosaic mutations

LIST OF PAPERS

This thesis is based on the following studies, referred to in the text by their Roman numerals (**I-VI**).

I. Kanter-Smoler G, Fritzell K, **Rohlin A**, Engwall Y, Hallberg B, Bergman A, Meuller J, Grönberg H, Karlsson P, Björk J, Nordling M. Clinical characterization and the mutation spectrum in Swedish adenomatous polyposis families. *BMC Med. 2008 Apr 24;6:10*.

II. **Rohlin A**, Engwall Y, Fritzell K, Göransson K, Bergsten A, Einbeigi Z, Nilbert M, Karlsson P, Björk J, Nordling M. Inactivation of promoter 1B of APC causes partial gene silencing: evidence for a significant role of the promoter in regulation and causative of familial adenomatous polyposis. *Oncogene. 2011 Dec 15;30(50):4977-89*.

III. **Rohlin A**, Wernersson J, Engwall Y, Wiklund, L, Björk J, Nordling M Parallel sequencing used in detection of mosaic mutations: comparison with four diagnostic DNA screening techniques. *Hum Mutat Jan 30:1012-1020, 2009*.

IV. **Rohlin A**, Eiengård F, Lundstam U, Zagoras T, Nilsson S, Edsjö A, Pedersen J, Svensson JH, Skullman S, Karlsson GB, Nordling M. Whole exome sequencing in hereditary colorectal cancer syndromes. Identification of causative mutations and contributing variants. *Submitted Manuscript*

V. **Rohlin A**, Zagoras T, Nilsson S, Lundstam U, Wahlström J, Hultén L, Martinsson T, Karlsson GB, Nordling M. A mutation in POLE predisposing to a multi-tumor phenotype. *Int J Oncol. 2014 Jul;45(1):77-81*.

VI. **Rohlin A**, Rambech E, Kvist A, Eiengård F, Wernersson J, Lundstam U, Zagoras T, Törngren T, Borg Å, Björk, J, Nilbert M, Nordling M. A validated multigene panel for colorectal cancer *Manuscript*

ABBREVIATIONS

AFAP	Attenuated Familial Adenomaotus Polyposis
APC	Adenomatous Polyposis Coli
BMPR1A	Bone morphogenetic protein receptor type 1A
BRAF	v-raf murine sarcoma viral oncogene homologue B1
bp	base pair
cDNA	complementary DNA
CIN	Chromosome instability
CNV	copy number variant
COSMIC	Cataloge Of Somatic Mutations In Cancer
CpG	cytosine-guanine dinucleotide
ddNTP	dideoxynucleotides
DHPLC	Denaturating high-pressure liquid chromatography
DNA	deoxyribonucleic acid
ds DNA	double stranded DNA
EMD	exonuclease domain mutant
FAP	Familial Adenomatous Polyposis
GREM1	Gremlin 1
HGMD	Human Genome Mutation Database
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
InSiGHT	The International Society for Gastrointestinal Hereditary Tumors
IHC	Immunohistochemistry
KRAS	Kirsten rat sarcoma viral oncogene homologue
LOH	loss of heterozygosity
LOVD	Leiden open source variation database
MAP	MUTYH Associated Polyposis
MLH	Mut L homologue
MLPA	multiplex ligation-dependent probe amplification
MMR	Miss-Match repair
mRNA	messenger RNA
MSH	Mut S homologue
MSI	micro satellite instable
MSI-H	micro-satellite instability high
MSS	micro satellite stable
MUTYH	Mut Y homologue
PCR	polymerase chain reaction
POLD1	DNA polymerase delta catalytic subunit
POLE	DNA polymerase epsilon
PPAP	Polymerase Proofreading Associated Polyposis
PTEN	Phosphatase and tensin homologue
RNA	ribonucleic acid
rRNA	ribosomal RNA
RT-PCR	reverse transcriptase PCR
SMAD	Mothers against decapentaplegic homologue
SNP	single nucleotide polymorphism
SNV	single nucleotide variant
STK11	Serine/threonine kinase 11
SV	structural variations
TGFβ	Transforming growth factor beta
TP53	Tumor protein 53
TSG	tumor suppressor gene
UCSC	University of California, Santa Cruz
UTR	Untranslatedregion
wt	wild type

INTRODUCTION

Basic Genetics

DNA and Genes

In humans the genome consists of DNA (deoxyribonucleic acid) and can be found in the nucleus and the mitochondria. The DNA is built from four different nucleotides; adenine (A), cytosine (C), guanine (G) and thymine (T) and linked together by covalent phosphodiester bonds that join the 5´carbon of the deoxyribose group to the 3´carbon of the next nucleotide. The DNA is formed as a double helix, held together by complementary hydrogen bonds between A-T and C-G base pairs and was first described by Watson and Crick in 1953 [1]. The human genome consists of approximately 3 billion bases (bp) organized into 23 chromosome pairs. The usages of these bases in in different combinations make up the genetic code.

A gene can be described as a region of genomic sequence, which is associated with regulatory regions, transcribed regions, and/or other functional sequence regions that contribute to phenotype or function as described in the official guideline for Human Gene Nomenclature. The exact number of genes is still not known but there are around 23,000 genes, which make up 1% of the total genome. In the classical view of a gene, it includes exon, introns and a promoter region. The promoter constitutes the regulatory region in the 5´end of the gene, where transcription factor binds and direct the transcription. Regions located far away from the gene including enhancer, silencers and insulator elements can also affect transcription. Genes are also expressed at different rates and in different tissues and can also be subjected to go through alternative splicing that further influence the complexity and diversity. There are also non-coding RNA (ncRNA) and conserved regions outside the genes that can perform function, which challenge the concept of a gene.

The central dogma

The expression and translation of genes is often referred to as the central dogma of molecular biology (Figure 1). This process is initiated by transcription of the DNA into a pre-mRNA followed by splicing of the pre-RNA into the mature messenger RNA and post-transcriptional processing. The mRNA migrates from the nucleus to the cytoplasm where it serves as a template in the translation process from RNA to protein on the ribosomes. The protein consists of different amino acids translated from a codon of three nucleotides in the mRNA named the genetic code. The protein undergoes different post translational modifications and folds up into a unique three-dimensional configuration to yield the final active protein.

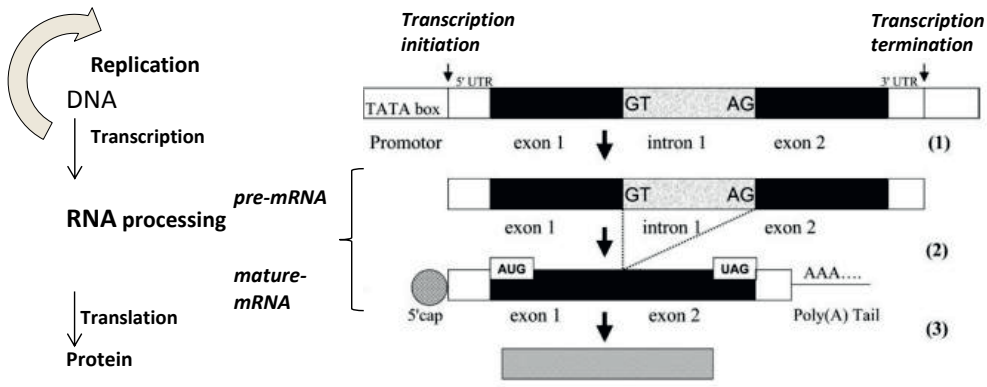

Figure 1. *The central dogma in molecular biology. The gene is transcribed (1) into primary RNA with coding exons and non-coding introns. The primary RNA is subjected to splicing (2) of the introns to yield the mature mRNA that contains exons and flanking sequences of untranslated regions (UTR) and includes 5´capping and 3´polyadenylation. Translation (3) of the mature mRNA into the polypeptide starts at the AUG codon and ends at the stop codon UAG. U, Uracil in RNA corresponds to T, Thymine in DNA (Reprinted and modified from Knoers and Monnens 2006).*

Splicing

The splicing process involves the removal of introns and rejoining of exons. RNA splicing requires a donor site (5' end of the intron), a branch site (near the 3' end of the intron) and an acceptor site (3' end of the intron). The splice donor-site includes almost an invariant GU sequence and the splice acceptor site a highly conserved AG sequence (Figure 1). The splicing process is mediated by the spliceosome complex consisting of small nuclear RNA and more than 50 proteins [2]. Alternative splicing is the process where the RNA can be reconnected in multiple ways resulting in different isoforms.

Epigenetics

Epigenetics refers to heritable changes in gene expression and does not involve changes to the underlying DNA sequence; a change in phenotype without any genotype change. At least three systems; DNA methylation, histone modifications and non-coding RNA - associated gene silencing are considered to initiate and sustain epigenetic changes.

The cytosine at CpG sites can be modified by methylation. This is common at CpG sites in repetitive sequences throughout the genome. CpG sites are also common in promoter regions and in the first exon of a gene and these sites are by default unmethylated. Cancer is characterized by genome wide hypo methylation together with gene specific hypo- or hyper methylation. Tumor suppressor genes are often inactivated

trough hypermethylation of promoter CpG islands. Histone modifications include acetylation, methylations, glycosylation or ubiquitination and combinations of these modifications constitute the histone code.

Mendelian Inheritance

Genes are inherited in two copies one from each parent. A gene may have different alleles but only two of them will appear in the same individual, so the genotype of an individual is represented by two alleles of each gene. Disease with monogenetic inheritance are caused by single locus variations in the genome, a dominant inheritance of only one allele in the locus decide the phenotype and a recessive inheritance require two alleles that signify the specific character in order to have an effect on the individual. The heterozygote genotype harbors a difference in the DNA sequence between the two inherited alleles. The genotype in recessive inheritance can be homozygous at a locus, where the DNA sequences of the two alleles are identical. It could also be a compound heterozygote where there are two different heterozygote mutations, one on each allele.

Linkage

Linkage means that two loci that are located adjacent to each other have a greater chance of being inherited together during meiosis than could be attributed by chance. At meiosis the crossing over between maternal and paternal chromosomes will produce recombinant chromosomes. If two loci are closely located on the chromosome it is less likely that a recombination occur between them. The recombination fraction (Θ), which is defined as the probability of a recombination separating two loci, can be used as a measurement of distance. The lower the recombination fraction, the stronger the linkage of the two loci. The genetic unit centiMorgan (cM) is often used in linkage maps and 1cM represents the distance between two loci that are on the average recombined once in 100 meiosis.

When DNA can be collected from several individuals in a family, preferable both affected and unaffected, a genome-scan with polymorphic markers either microsatellite or SNP markers can be used to identify genomic regions that segregated with the affected individuals in the family. Genotype data from individuals and marker information are used to estimate the likelihood of a marker being linked to the disease locus. The likelihood of linkage divided by the likelihood of no linkage for a specific marker quantifies linkage. The base 10 logarithm of this likelihood ratio is defined as the LOD score (logarithm of odds ratio), where a LOD score thus is a measure of linkage. Linkage analysis is a useful tool in trying to identify genes that are associated with disease in combination with exome sequencing.

In paper IV and V we used Affymetrix SNP array 6.0 for genotyping of affected and unaffected family members. The linkage analysis was done with a parametric linkage model (see method section). This model assumes a dominant inheritance and is most suited for high penetrance and rare diseases. A LOD score threshold can be set,

defining a small number of regions where the disease-causing mutation might be found. These regions can provide a start point in selecting variants in the exome analysis, which was done in paper IV in family C.

Variations in the genome; polymorphism and mutations

SNVs, small insertion/deletion variants

DNA-sequence variations can be of different kind, Single Nucleotide Variants (SNVs), insertions and deletions. SNVs variations in which one nucleotide differ between individuals are the most common ones. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNVs in the human genome. A SNV has normally two alleles, but three or four do exist. When SNVs are located in the coding part of a gene, they may affect the amino acid. A variant that have a profound effect on the amino acid is called non-synomomous otherwise they are called synonomous. Non-synonomous variants can be divided into missense and nonsense. Missense variants results in a different amino acid and nonsense in a stop codon. Mutations are in this thesis defined as DNA variants that are defined as pathogenic, whereas polymorphisms are defined as non-pathogenic. In a population, variations can be assigned a minor allele frequency (MAF), which means the lowest allele frequency at a specific locus in a particular population. A polymorphic variant can also be defined as variant occurring in more than 1% of the population. There are variations between human populations, a variant that is common in one geographic or ethnic groups might be rare in another. A nonsense variant is often a mutation as it results in a premature stop codon, called a loss of function variant (LoF). However sometimes, but rarely a read through the stop codon can result in a functional protein product or a LoF that is not harmful [3]. Small insertion and deletion of one to several bases are often pathogenic if they occur in the coding region and cause a frameshift in the reading frame and eventually a downstream stop codon. In frame deletions and insertions are more difficult to interpret the effect of.

Missense variant prediction and classification

The disease-association of a missense variant is often more difficult to interpret, because an amino-acid substitution can affect the biological function of the protein in a number of different ways. It may disrupt catalytic residues or ligand-binding pockets and/or lead to alterations in structure, folding or stability of the protein [4] Several *in silico* protein predication programs exists that predict the outcome of a missense change, these can be divided into at least two types, conservation based predictor and trained classifiers. Conservation based predictors like SIFT assume that functional substitutions occur at sites that are evolutionary conserved and uses protein homology (multiple sequence alignment) across species to calculate position specific scores. Some of these methods, e.g. Polyphen-2, also include biochemical structural data like the three-dimensional structure of the protein. They calculate the effect in the surrounding residues by considering changes in size, polarity, protein stability and electrostatics,

which can significantly improve the prediction of deleteriousness. Polyphen-2 and MutationTaster combine multiple sequence alignments and structural information and in addition they are trained to differentiate as set of true deleterious and benign variants and are therefore called trained classifiers [5]. There are also programs that make predications by combining the output from other programs for example Condel (consensus deleteriousness score of missense mutations) [6] and PON-P (Pathogenic or Not Pipeline) [7] which uses a combination of five different predictors in order to assess the deleteriousness of variants.

Nucleotide-based predictors can be used for coding and non-coding DNA, they do base their prediction on evolutionary conservation and estimate observed rate of evolutionary changes and compare this with expected rates for neutral positions, sites with fewer substitutions receive higher scores. A method like this is phastCons [8] which uses a model in which also the score of neighboring nucleotides is taken into account whereas others consider each position independently like phyloP [9].

New methods for *in silico* protein prediction are constantly evolving. A newly published method for estimating the relative pathogenicity is Combined Annotation-Dependent Depletion (CADD) which is a method to measure deleteriousness by contrast the annotation of fixed or almost fixed derived alleles with those of simulated variants. In this method a combination of several parameters are used including, allelic diversity, annotation and functionality, pathogenicity, disease severity, experimentally measured regulatory effects and complex trait associations and highly ranked know pathogenic variants within individual genomes. Variants that are more likely to be simulated, not observed, are more likely to have a deleterious effect. This is measured in a Phred-like scale C-score, where a score of 10 represent the 10% most deleterious substitutions that can be done to the human genome and a score of 20 represents the 1% most deleterious variants. A cutoff around 15 is recommended as a guideline for deleteriousness [10]. This method is used in paper VI.

Databases

Databases in which different consortiums have made data publically available are an additional tool in the evaluation and classification of variants. These databases are constantly growing, some of the most common used ones are: The Single Nucleotide Polymorphism database (dbSNP http://www.ncbi.nlm.nih.gov/SNP/), the 1000 Genomes project (http://www.1000genomes.org), the exome variant server(http://evs.gs.washington.edu) with a collection of 6,503 exome sequences, the catalogue of Somatic Mutations in Cancer (COSMIC) (http://cancer.sanger.ac.uk/), Human Genome Mutation Database (HGMD http: //www.hgmd.cf.ac.uk) and the locus specific Leiden Open source Variation Database (LOVD). In addition in-house dataset or databases are very useful for information regarding local common variants.

Guidelines for classifying variants

Guidelines for classifying variants in the mis-match repair genes have been published from the International Agency for Research on Cancer (IARC) [11] and InSiGHT (International Society for Gastrointestinal Hereditary Tumors) [12] according to a five-class system: 1 = Benign, 2 = Likely benign, 3 = Variant Of Unknown clinical Significance (VOUS), 4 = Likely pathogenic and 5 = Pathogenic.

Splice effecting variants

When changes (substitutions, insertion, deletions) occur in donor sites or acceptor sites the splicing of the exon can be affected and often these changes are mutations. Nucleotide changes outside the donor and acceptor –sites can also be mutations, intronic splice elements and changes in the number of nucleotides between the branch point and the nearest 3´acceptor site can affect splice site selection. Variations can create cryptic splice-sites, which results in exons that loose a part of the exon or gain a novel part from the intron, which can manifest as a truncation of the final protein. [13,14] (Figure 2).

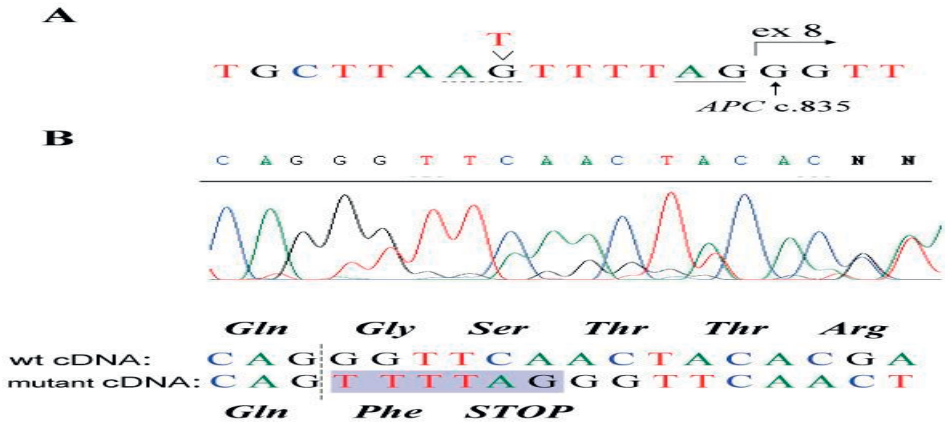

Figure 2A) *Genomic sequence of a patient with a c.835-7T > G mutation. The new splice site generated by the T > G substitution is indicated with a dashed line, the wildtype acceptor-splice site is underlined, and the regular start of exon 8 is indicated with an arrow.* ***B)*** *cDNA sequence covering the exon 7–8 boundary, indicated with a dashed line. Shown below the sequence diagram is the interpretation of the sequence reflecting the two mRNA species present in the sample. The insertion of 6 bp owing to the introduction of a new splice site in the mutant allele is shown as a shaded area. Predicted amino-acid sequence of translation products are shown above and below the respective cDNA sequence (Reprinted from [15]).*

Structural variants

Structural variants (SV)s was originally defined as, deletions, insertions and inversions greater than 1 kb in size. With gained knowledge of the human genome the spectrum of structural variants (SVs) has broaden to include genomic rearrangements that affect >50 bp of sequence and up to large-scale aberrations involving the loss or gain of whole chromosomes, numeric aberrations, loss or gain of parts of chromosomes called segmental or structural aberrations and translocations and rearrangements of parts between non-homologous chromosomes. Numerous classes of SVs exists and include deletions, tandem duplications, novel insertions, inversions, translocations and mobile elements [16].

In general SVs encompassing deletion or insertion of several exons of a gene, a whole gene or gene region, in known disease-causing genes or gene regions, are pathogenic mutations. Concerning more unexplored genes or gene regions, higher caution has to be taken as copy number variations (CNVs), a large category of structural variants (typically greater than 1 kb and less than five Mb) can also display differences in normal population [17]

Mosaic variants

Mosaicism is defined as the presence of two or more populations of cells in an individual and is developed from postzygotic mutations. Prezygotic mutational events can also result in a parent who is mosaic (gonadal mosaicism), and the mutation might be inherited in the zygot and in all cells of the developing offspring. The variant can be present in one or several of the germlayers, organ and organ systems. The timing and tissue of origin will have consequences for whether or not the mutation will be transmitted to the offspring and to following generation [18]. Mosaic variants are common in hereditary disorders with a relative high frequency of *de novo* mutations, >30% of NF2 patients with new mutations are estimated to be mosaic [19] and 20% of FAP patient [20-22]. The advances in massively parallel sequencing have made it easier to detect these mosaic mutations. The individual molecule sequencing allows absolute quantification of the mutant allele. Several other genes harboring mosaic mutations have recently been presented like PTEN [23,24], TP53 [25] and PPM1D [26].

Variants in regulatory regions

Variants in the promoter region can include larger structural variants including the whole promoter region or a part of it, they can also be small abbreviations which can have effect if they perturb transcription binding site and/or CpG sites. Enhancers are sequence elements that bind activators, they are linked in cis with a promoter and stimulate its activity. They are typically a few hundred base pairs long and include binding sites for transcription factors. Enhancer can regulate multiple neighboring genes far away and even interaction between enhancers and promoters on different chromosomes have been observed. The mechanisms involved in the enhancer-

promoter interaction are poorly understood, but are now thought to include biochemical compatibility, spatial architecture, insulator element and the effect of local chromatin composition. Insulator elements are elements that can prevent the activation of promoters by an enhancer, when placed between them [27]. Structural variants ranging from deletions, tandem duplications and or inversions have been found to re-positioning genes next to super-enhancers. Super-enhancers have recently been identified as regions with a concentration of activators and transcription factor binding, which stimulate higher transcription than normal enhancers [28]. Mutations in regulatory regions are identified in paper II and IV. In paper II we identified a mutation (deletion) including half of promoter 1B of the *APC* gene and in paper IV we identified a mutation (duplication) of a region including an enhancer element near the *GREM1* gene.

Loss of function variants (LoF)s

In 2012 MacArthur et al [3] reported a list of 1,285 high confidence loss of function (LoF) variants by analyzing 1000 Genome samples data, were 32% were predicted to affect functional proteins. They estimated that there were around 100 LoFs per human genome in healthy individuals, and around 20 of these in a homozygous state. Lately further analysis have been focusing on rare germline variants (<1%), not pathogenic for the disease or phenotype investigated and present in any human genome. Guidelines are being proposed distinguishing disease-causing sequence variants from functional variants that do not cause disease and are present in any human genome [29]

Genetic analyses in hereditary cancer disease

Genetic analysis of hereditary cancers is primarily performed by studies of DNA from blood, but tissue samples fresh or formalin fixed paraffin embedded (FFPE) can be used as well. In some cases when a mosaic mutation is suspected, tissue samples from different germ layers (endoderm, ectoderm and mesoderm) are preferred as a mosaic mutation can be present in one or the other of the germ layers. Tumor samples can also be analyzed if they are available. Mutations that predict to result in a truncation of the protein, nonsense mutations, short deletions/insertions associated with a frame shift, mutations involving position +/- 1 and +/- 2 (related to the exon) within splice junctions and large rearrangement, are likely to impair the protein function and are usually classified as disease causing without any additional information. However, in cases with mutations involving nucleotides outside the highly conserved splice junction positions, RNA has to be collected as well in order to analyze for splice effects on the transcription level. Missense variants are more difficult to interpret, synonomous variations are in general classified as likely benign, not disease causing, as long as they are not predicted to have any splice effect. Non-synonomous variants have to be very carefully interpreted, segregation analyses in combination with documented functional effects are preferred in order to assess their pathogenicity. Databases of normal variants as well as the use of local normal controls are also important tools used to classify these variants correctly. General guidelines on genetic and mutation nomenclature are

necessary for a correct interpretation of a genetic analysis. However, the evolving guidelines may sometimes be problematic and confusing when well established mutation annotations suddenly become incorrect according to novel guidelines. The current recommendations are provided by the Human Genome Variation Society (HGVS) [30,31]

Cancer Genetics

Cancer

Cancer is a genetic disease, all cancers arise as a result of several somatically acquired changes in the DNA of a cancer cell or rarely as an inherited predisposition. Cancer is not one disease more than hundred different types exists and over the last decade huge sequencing efforts have revealed the genomic landscape of many common forms of cancers. For most cancer types the genomic landscape consists of small numbers of "mountains" which are genes that are altered in high percentages in tumors and a much larger number of "hills" that are genes altered infrequently. This new view of cancer is consistent with the idea that a large number of mutations, each associated with a small fitness advantage, drive tumor progression. It is the hills and not the mountains that dominate the cancer genome landscape [32]. However, the hills represent alterations in much smaller number of cell signaling pathways and these pathways rather than single genes, drive the course of tumorigenesis. 12 pathways have been identified that regulate three core processes: cell fate, cell survival and genome maintenance. Not all somatic abnormalities in a cancer genome have been involved in the development of tumors or are necessary for the cancer progression and therefore the concept of driver and passenger mutations is used. A driver mutation confers a growth advantage and has been positively selected in the micro environment of the tissue in which the cancer arise. A typical colorectal tumor contains about 80 mutations, around, 2-8 of these are driver mutations and the remaining mutations are passengers [33,34]. Historically there are two major groups of genes frequently altered in cancer. These are oncogenes and tumor suppressor genes (TSG).

Oncogenes

Oncogenes are altered versions of normal proto-oncogenes. These genes normally have cell proliferating functions involving regulation and progression of the cell cycle, cell division and differentiation. Mutations in these genes result in a gain of function, which means an excessively or inappropriate activation (oncogene). Alteration of one allele of an oncogene is sufficient to affect the phenotype of the cell.

Tumor suppressor genes

Tumor suppressor genes (TSG) are inhibiting uncontrolled cell growth, mutation in these genes result in loss of function and both of the alleles are need to be inactivated in order to affect the phenotype. The theory behind is explained in Knudson´s two-hit

hypothesis. This theory states that two hits are needed for a TSG to be inactivated, and is based on retinoblastoma development (a tumor in the eye) [35]. The first hit can either be inherited like a germ-line mutation or acquired somatically, the second hit is always a somatic mutation. A individual that inherit a TSG mutation, which can be a point mutation, small or large deletion, insertion duplication or hypermethylation, will carry the mutation in all cells and only one further somatic hit is necessary in any of the cells in a relevant tissue to get a loss of function of the protein. In the tumor one allele of a TSG is often but not always lost as a large deletion of the chromosomal region. Deletions like this are often discovered in tumor cells by loss of heterozygosity (LOH) studies, which can be used in order to identify novel tumor suppressor genes.

TSGs can be divided into gatekeepers and caretakers[36]. The gatekeepers are directly regulating the growth of tumors, maintaining a constant cell number by inhibiting growth and promoting apoptosis. Both the maternal and the paternal alleles need to be inactivated for tumor initiation, The *APC, VHL*, *NF1, RB* and *TP53* genes, associated with dominant familial cancer syndromes, are gatekeepers. Caretakers or DNA stability genes, promote tumor growth more indirectly which leads to genomic instability and an increased mutation rate in other genes. The mis-match repair (MMR) genes involved in Lynch syndrome and the *MUTYH* gene are examples of caretaker genes involved in familial colorectal cancer syndromes.

New insights and classification of oncogenes and tumor suppressor genes

The divergence of oncogenes and tumor suppressor genes are now more based on mutation patterns. An oncogene has been defined as a gene where > 20% of the mutations are at recurrent sites and are missense leading to amino-acid substitutions. A tumor suppressor gene is defined as a gene where >20% of the mutations in the gene are inactivating. Genes can also be both an oncogene and a tumor suppressor gene in different context, which for example is demonstrated by the *NOTCH1* gene. In lymphomas and leukemias, mutations in this gene are often recurrent missense mutations, where as in squamous cell carcinoma these mutations are often none recurrent and inactivating [33]. The *RET* gene is an oncogene in medullary thyroid carcionoma [37], but aberrant methylation of *RET* and inactivating mutations suggest that RET can function as a tumor suppressor gene in colon [38]. The knowledge that the same gene can function in opposite ways in different celltypes is important for understanding different cell-signaling pathways.

There is also a shift considering mutations that give rise to premature truncation of protein translation as it e.g. has been shown for the p53-inducible phosphates encoding gene, PPM1D, in which truncating mutations have activating oncogenic activity [26].

Colorectal polyps

Colorectal polyps are growth that project from the lining of the colon or rectum. They are seldom symptomatic, but their significance lies in their potential to form malignant

transformation. Histologically they are divided into hamartomatous, serrated and adenomatous polyps. Adenomas arise from the glandular epithelium and are characterized by dysplastic morphology and altered differentiation of the epithelial cells in the lesion. Small adenomas often have a tubular growth pattern whereas larger more often have a villous growth pattern, and are classified as advanced adenomas. Hamartomatous polyps, in e.g. juvenile polyposis (JP) have an expanded mesenchymal stroma with pronounced inflammatory infiltrate that consists primarily of lymphocytes and plasma cells. They show structural epithelial abnormalities at the level of crypt and architecture with an uncontrolled formation of new crypts and increased cellular proliferation, but the epithelial cells themselves show normal maturation and no dysplasia like in adenomas [39,40] Traditionally serrated adenomas and sessile serrated adenomas are related to hyperplastic polyps, however hyperplastic polys are considered benign whereas the sessile serrated adenoma and serrated adenoma are precancerous lesions.

Pathways to colorectal cancer

In 1990 Fearon and Vogelstein proposed a multistep genetic model, where the accumulation of multiple genetic mutations lead to a stepwise progression from normal to dysplastic epithelium in the colon [41]. Colorectal cancer was believed to progress through an adenoma carcinoma sequence that still might be true for the majority of CRCs that arise from premalignant adenomas including familial CRC syndromes. In the Vogelstein model APC/β-catenin mutations serve as the initiating step followed by RAS/RAF mutations and loss of p53 function at a later stage (Figure 3). Lately however the complexity reveals that epigenetic variations and non-coding RNAs are also important and the timing and combination of genetic and epigenetic events rather than the increased accumulation of genetic mutations appear to result in activation of district pathways that give cancer cells a selective disadvantage [42].

Most of the tumors in Lynch syndrome arise through conventional adenomas and by the traditional adenoma carcinoma sequencing of events. In Lynch syndrome activation mutations in *CTNNB1*, especially in exon 3, can be found in a proportion of tumors that do not harbor *APC* mutations [43].

Three major pathways leading to CRC, where originally described, chromosome instability pathway (CIN), Microsatellite instability pathway (MSI) and the CpG island methylation pathway (CIMP). Over the past few years however, new information has led to a classification that are more based on the genomic changes discovered in huge sequencing projects. In 2012 the Cancer genome atlas network published somatic alterations in 276 colon cancer samples found by exome sequencing, DNA copy number variation analysis, promoter methylation analyses, mRNA and micro-RNA expression analyses. Through these studies much have been learned about the heterogeneity of CRC tumors on the molecular level which can be used for guidance of the prognosis, response and treatment of CRC [44].

Chromosome Instability pathway (CIN)

The first and most common distinct molecular pathway is CIN. This pathway is defined by accumulation of numerical (aneuploidy) and/or structural chromosomal abnormalities and is characterized by frequent loss of heterozygosity (LOH) at tumor suppressor gene loci and by chromosomal rearrangement [45]. CIN tumors are also defined as non-hypermutated, they accumulate mutations in *APC* and *TP53* in much higher extent than the hypermutated tumors, they also accumulate mutations in *KRAS*, *PIC3CA*, *BRAF* and *SMAD4*. The CIN phenotype could result from defects in pathways that are involved in inaccurate chromosome segregation. The mitotic checkpoint (spindle assembly checkpoint) is the major cell cycle control mechanism that assures high fidelity of chromosome segregation by delaying the onset of anaphase until all pairs of duplicated chromatids are properly aligned. Defect in checkpoint signaling leads to mis-segregation and aneuploidy. Mitotic arrest-deficient (MAD) and budding unhibited by benzimidazoles (BUB) are checkpoint sensors and signal transducers that control sister chromatid separation [42,46,47].

Microsatellite instability pathway (MSI)

Microsatellite instability is caused by dysfunction of the MMR genes leading to mismatches in the DNA that are not repaired, which leads to an accumulation of mutations and a hyper-mutated phenotype. Microsatellites are nucleotide repeat sequences of 1-6 bp in length that are prone to accumulation of mutations because of DNA polymerase slippage leading to framshifting mutations which could cause protein truncation if they occur in coding regions. In the wild-type cell this is corrected by proteins encoded by the mismatch repair genes (MMR). Most of the microsatellites are found in noncoding regions, but some genes e.g. the TGF-β receptor type II and the IGF II receptor harbor microsatellites and are particularly prone to mutations in Lynch syndrome associated CRC. Microsatellite instability due to mutation in MMR genes (*MLH1, MSH2, MSH6* and *PMS2*) is the hallmark of tumors in Lynch syndrome. In diagnostics at least five markers are tested for microsatellite instability. If 30% or more show instability, the tumors are classified as MSI-High (MSI-H). MSI-H tumors are also found in about 15% of sporadic CRC caused by epigenetic silencing due to hypermethylation of MLH1 in both alleles [43,48].

Recently advances in high-throughput sequencing of tumors revealed new mutations and refined classification will probably emerge. The molecular characterization of CRC tumors in the TCGA project found that among hyper-mutated tumors approximately 75% were MSI-H with hypermethylation of the MLH1 promoter resulting in *MLH1* silencing. However, approximately 25% of tumors were found to be MSS with somatic mutations in the MMR genes and *POLE* (DNA polymerase ε). These tumors were shown to have an even higher mutation rate and are classified as having an ultramutator phenotype [44].

The CpG island methylation pathway (CIMP)

CRC tumors can also be classified based on methylation of CpG islands. Most of sporadic tumors have a widespread hypermethylated phenotype and can be classified as CIMP positive. There are tumors that have fewer methylated CpG islands and also show lower level of methylation at individual loci, these are classified as CIMP-low [49,50]. These tumors can further be divided into different subtypes related to harboring *BRAF* and *KRAS* mutations. [51-53].

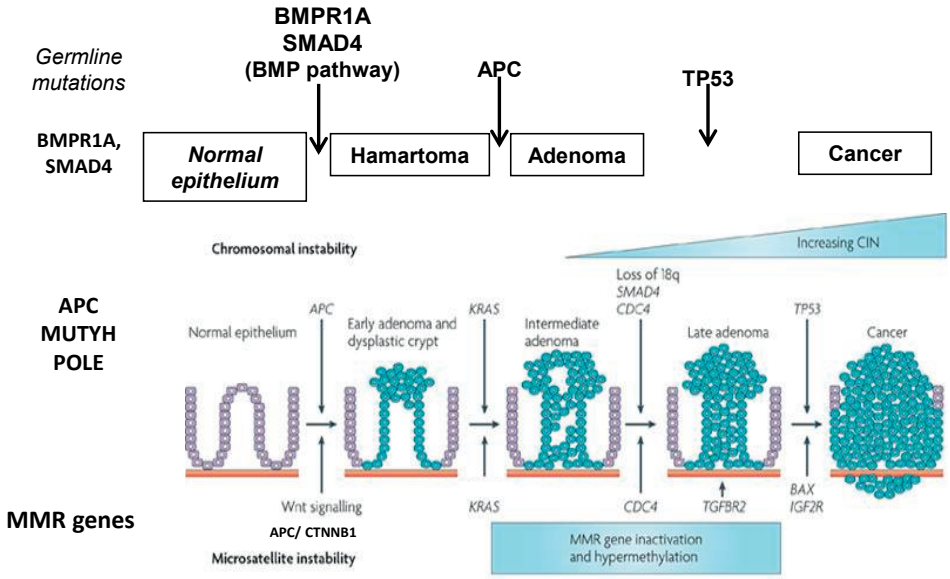

Figure 3 *A schematic simplified overview of adenoma to carcinoma progression, involving different germline initiating mutations and the genes subsequently mutated in the CIN and MSI respective pathways. (Reprinted and modified by permission from Macmillan Publishers Ltd: Nature Reviews Cancer][54], © (2009).*

Hereditary colorectal cancer

Colorectal cancer is the third most prevalent cancer [55] and the second most common cause of cancer mortality in the world [56]. Genetics has a key role in predisposition to CRC and kindred and twin studies have estimated that around one third of all CRC cases are an inherited form of the disease [57]. High penetrant mutations in known CRC predisposing genes explain only about 5-6% of the cases (Figure 4). All of these syndromes are based on clinical and pathological findings, but recently also genetic characterizing of the syndromes have been considered and used in the classification of the syndromes.

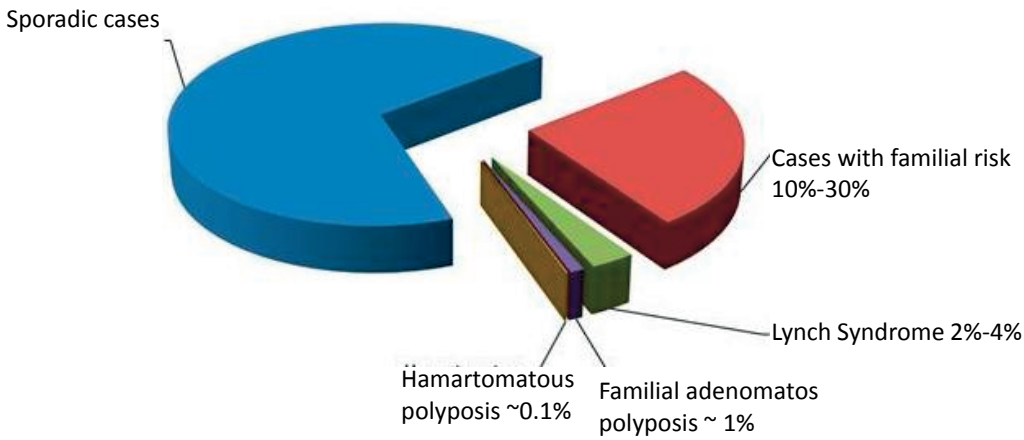

Figure 4. *The fraction of colon cancer cases that arise in various family risk settings (Reprinted and modified with permission from Elsevier: Gastroenterology [58] © 2000).*

Familial Adenomatous Polyposis (FAP)

FAP account for around 1% of CRC cases and is the second most common inherited CRC syndrome with a prevalence of 1 in 10,000-30,000 individuals. Characteristic features of FAP include hundreds to thousands of colonic adenomas beginning in early adolescence or in childhood, mostly in the distal colon which inescapable lead to CRC in untreated individuals. Generally cancers start to develop a decade after appearance of polyps. The average age of CRC diagnosis if untreated is 39 years; 7% develop CRC by age 21 and 95% by age 50. Other extra-colonic features include; fundic gland polyps in 90% of affected individuals, duodenal and periampullary polyps in more than 50%, duodenal cancer is also the second most common malignancy in FAP. Duodenal polyps are classified according to a scale based on polyp number, size, histology and severity of dysplasia which is referred to as Spigelman´s classification. Individuals with FAP also carry a risk of small bowel polyps.

Extra-colonic manifestations also occur in FAP, they are rarely malignant and include congenital hyperthrophy of the retinal pigment epithelium (CHRPE), osteomas, epidermoid cysts, fibromas, dental abnormalities and desmoids. Desmoids are soft-tissue tumors in the mesentery abdominal wall. These tumors are benign, but by progressive enlargement and by the consequent pressure they cause on gastrointestinal or urinary tract and local nervous system they can be life threatening and cause severe morbidity as well as mortality. Desmoids occur in around 8% of men and 13% of women with FAP. Other extra-colonic cancers include thyroid, bile duct, liver (hepatoblastoma) and central nervous system (cerebellar medulloblastoma). The association of colonic adenomas together with lesions outside the colon is also called Gardners syndrome [56,59-61].

Attenuated FAP (AFAP)

Attenuated FAP is a less aggressive variant of FAP characterized by fewer adenomas, usually around 10-100. Patients have a later age of adenoma appearance and most of the adenomas are found in the proximal colon. Even though they have fewer polyps patients have an increased risk of cancer, generally occurring 10-15 years later than in FAP. As in FAP duodenal and gastric fundic gland polyps are common, but extra colonic manifestations as those found in FAP are rare. Attenuated FAP can mimic typical FAP, MUTYH Associated Polyposis (MAP) or even sporadic polyp development. Attenuated FAP and MAP respectively account for 10 % to 20 % of individuals with 10 to 100 polyps [62,63].

The APC gene and mutations

FAP is autosomal dominantly inherited and is caused by germline mutations in the *APC* gene. The *APC* gene is a tumor suppressor gene and FAP patients inherit one mutant-*APC* allele followed by a somatic mutation in the remaining wt allele which initiates tumorigenesis. More than 1,000 mutations of the *APC* gene have been described. In classical FAP almost 100 % of the disease-causing mutations can today be found and 95 % of the mutations cause a truncation of the protein [15]. In AFAP only around 20-30 % of the disease-causing mutations can be identified in the *APC* gene. New or *de novo APC* mutations are responsible for approximately 25% of FAP cases and around 20 % of *de novo* cases have somatic mosaicism [20-22].

The *APC* gene is a tumor suppressor gene located on chromosome 15q21. The main transcribed coding region consists of 15 exons (where exon 15 encompass around 75% of the sequence) and encodes a protein product of 2,843 amino acids [64]. Two promoter regions have been identified, promoter 1A and promoter 1B and several alternative transcripts, which could also be tissue specific expressed exist. The *APC* gene is expressed in all tissues at various levels. The differentially alternative transcripts involve mainly the 5´part of the gene, exon 9 and 10A and some isoforms are tissue specific expressed, for example a brain specific exon (BS) exists. The alternative splicing mechanism involving exon 9 with removal of codon 312 to 412 produce a shorter *APC* isoform, both isoforms are present in normal tissue.

At least five transcriptional start-site, with transcription from both promoter 1A and 1B, have been identified [65-67]. In paper III in this thesis the expression of three different transcripts are investigated. These are NM_0011275.1 (11,025 bp) which, represents the longest transcript and is transcribed from promoter 1B, transcript NM_001127510.1 (10,838bp) which contains an alternative in-frame exon (1A) compared with NM_001127511.1 and finally NM_000038.5 which is 10,740bp. NM_000038.5 and NM_001127510.1 both represent the same isoform, but differ by 98 bp in the 5'- UTR, NM_000038.5 is usually the main reference transcript used.

The APC protein

The APC protein is a multifunctional protein and apart from its main role in Wnt signaling it is also involved in cell adhesion and migration, organization of the cytoskeleton, spindle formation and chromosome segregation, cell cycle regulation and apoptosis. APC plays a central role in Wnt signaling, by regulating the degradation of β-catenin, by acting in the destruction complex together with axin, glycogen synthase kinase (GSK3) and casein kinase 1 (CK1) alpha. Formation of this complex targets β-catenin for Ser/Thr phosphorylation and recognition by an E3 ubiquitin-ligase for degradation. In the absence of a signal from an extracellular Wnt ligand or the presence of wt APC protein, β-catenin is degraded. In the presence of an extracellular Wnt ligand or absence of APC, β-catenin levels rise, it enters the nucleus and binds to T-cell factor (TCF)-family DNA binding proteins and activates Wnt-respons target genes[68-70].

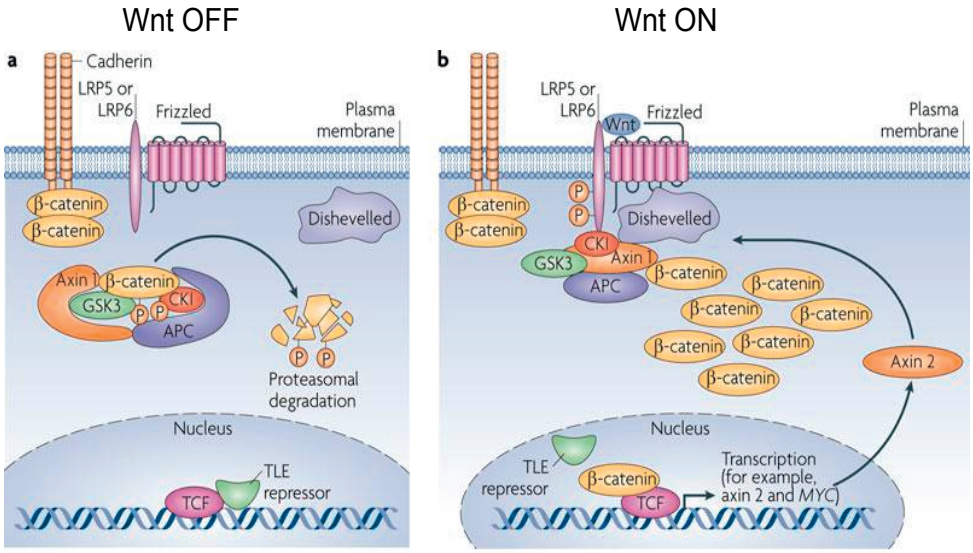

Figure 5. Wnt signaling pathway a) *In the absence of a signal, the destruction complex adenomatous polyposis coli (APC), axin 1, glycogen synthase kinase 3 (GSK3) and casein kinase 1 (CK1) binds and phosphorylates β-catenin, targeting it for destruction by the proteasome.* ***b)*** *The binding of a Wnt ligand to receptor or the absence of APC induces a change in conformation that results in disruption of the destruction complex. β-catenin can then accumulate and associate with the TCF proteins, dislodging the TLE repressors and hence promoting transcriptional activation of a programme of genes.(Reprinted by permission from Macmillan Publishers Ltd: Nature Reviews Molecular Cell Biology] [71], © (2010).*

The APC protein contains several protein interaction domains. At the N-terminal, APC contains an oligomerization domain allowing APC to form homo-dimers. Wild-type APC may form dimers with both wt and truncated mutant APC proteins. If the amount

of wt APC is reduced not only by mutated proteins, but also by dimerization of the remaining wt APC with mutant protein, a dominant negative effect may appear. The central part of the APC protein has a role in binding and degradation of β-catenin involving four 15 aa-repeat and seven 20 aa-repeat segments. In the C-terminal a microtubular-binding domain is located Figure 5 [59,72].

The just right signaling model

The region between codon 1250 and 1450 is referred to as the mutations cluster region. The first 20-amino acid repeat in the *APC* gene is located at the 5´-end of the mutation cluster region (MCR). The "just right signaling model" was proposed for the location of the first and second hit in the *APC* gene, in regards to the number of 20-aa repeats retained in the final protein. The ability of APC to regulate β-catenin activity appeared to be dependent on the number of 20-aa repeats present in the protein. Around two 20-aa repeats seemed to be associated with a suboptimal level of β-catenin signaling favoring tumor formation. In this way the first hit determined the type and location of the second hit. Germ-line mutations between the first and the second 20-aa repeats are associated with LOH of the *APC* locus and germline mutations before the first 20-aa repeat are associated with somatic mutations between the second and the third 20-aa repeat. Germline mutations after the second 20-aa repeat are associated with somatic mutations before the first 20-aa repeat [73,74]. Further studies have shown that this model also applies to extra-colonic lesions in FAP, but the combinations of mutations are different. In desmoids and upper gastrointestinal tumors, LOH is associated with germline *APC* mutations between the second and third 20-aa repeat and the optimum protein encode at total of three or four 20-aa repeats[75]. Some AFAP tumors have been found to acquire three hits at *APC*, this has particularly been reported in patients with exon 9 mutations [76].

Genotype phenotype correlations

There exist some correlations between the site of specific mutations and the clinical manifestation of the disease. Mutations contributing to classical FAP tend to occur between exon 5 and the 5´part of exon 15, where those associated with AFAP tend to cluster in the extreme 5´portion of the gene, in exon 9 most frequently in the alternative spliced region and in the 3´portion of exon 15. Mutations between codon 1250 and 1464 are associated with severe FAP. Mutations that cause CHRPE are associated with mutations that occur after exon 9. Patients with mutations between codon 1445 and 1587 can develop severe desmoid tumors [61,76,77] (Figure 6).

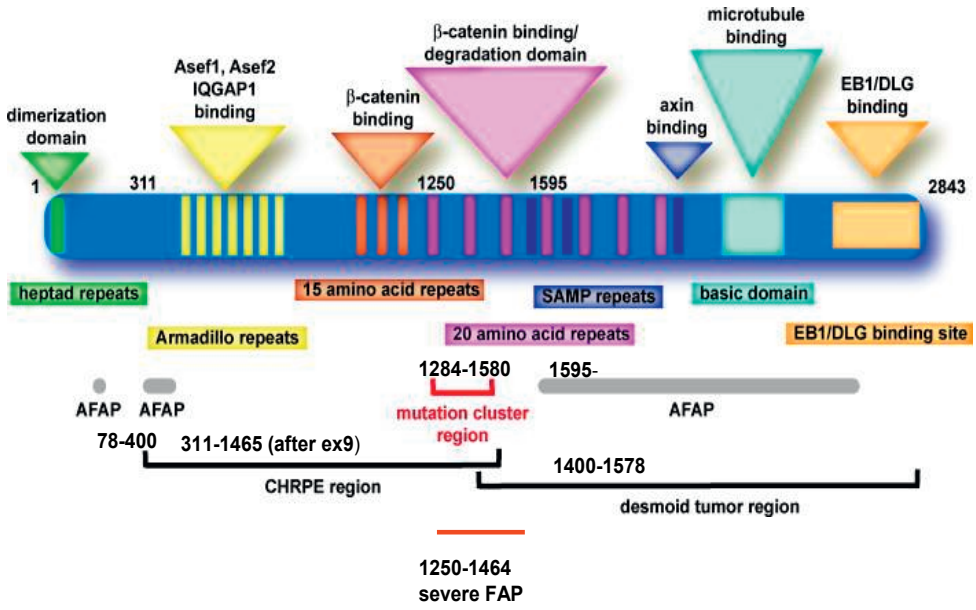

Figure 6. *The functional domains of the APC protein. Shown are the identified amino acid domains of the APC protein (rectangles) and the implicated functions of each domain (triangles). Also highlighted are particular disease phenotypes that appear to associate with mutations that truncate the APC protein in certain regions along with key codon positions along the protein (Reprinted and modified with permission from Elsevier: Mutation Research [77], © 2010).*

MUTYH associated polyposis

MUTYH associated polyposis (MAP) is characterized by the presence of around 30-100 adenomas mainly in the proximal colon and patients have an increased risk of CRC in their 4th or 5th decade of life. The colonic phenotype of MAP mimics AFAP, but although adenomatous polyps predominate in MAP, hyperplastic polys and/or sessile serrated adenomas/polys have been reported, which are not seen in AFAP. Gastric and duodenal polyps occur in around 11 % and 17 % respectively. Other FAP associated extra-colonic features such as osteomas, desmoids, CHRPE and thyroid cancer are not common, instead an excess of ovarian, bladder, skin sebaceous glad tumors and possibly breast cancer is observed, overlapping partly with the cancer spectrum of HNPCC [60,78-80].

MAP is inherited recessively with biallelic mutations in the *MUTYH* gene. Around 20%-30 % of *APC* mutations negative polyposis cases can be attributed to biallelic mutations in the *MUTYH* gene. The *MUTYH* gene is located on chromosome 1p32.1-p34.1 and the longest main transcribed transcript consists of 16 exons

(NM_001128425.1).Today around 300 variants have been identified in the *MUTYH* gene including 80 pathogenic mutations distributed throughout the gene. Various types of mutations have been reported including nonsense, small insertions/deletions, splice variants and missense mutations, which represent the majority of detected changes. The two most common mutations are missense mutations Tyr179Cys and Gly396Asp which represent 70% of the mutations found in European patients. There is a controversy regarding the CRC risk in individuals with mono-allelic mutations in the *MUTYH* gene. Three classes of mRNAs that include at least ten different transcripts are tissue-specific expressed with the occurrence of splicing events in the first and third exon of the gene [15,80].

MUTYH encodes a DNA glycosylase that is expressed both in the nucleus and the mitochondria. MUTYH glycosylase, is involved in base excision repair (BER), caused by oxidation. DNA oxidation arise from interaction with exogenous molecules or from the action of reactive oxygene species (ROS), results in G: C to T:A transversion mutations. MUTYH interacts with multiple replication and repair proteins there among the MSH2/MSH6 heterodimeric complex [81,82].

Hamartomatous polyposis syndromes

Hamartomatous Polyposis Syndrome (HPS) are characterized by the development of hamartoumatous polyps in the gastrointestinal tract (GI-tract). Hamartomas result from an abnormal formation of normal tissue, growing at the same rate as surrounding tissue. They are rare compared to neoplastic and hyperplastic polyps, but are the most common polyps in children. The hamartomatous polyps can vary in size and they have different histological structures, which makes it possible to distinguish between the different syndromes [83].

Peutz-Jegher Syndrome

Peutz-Jeghers Syndrome (PJS) is characterized by mucocutaneous melaotic pigmentation and hamartomatous polyps throughout the GI-tract and with gastrointestinal and extraintestinal cancer. There is a high rate of extra-colonic cancers including gastric, small bowel, pancreatic, breast, ovarian, lung, cervical and uterine/testicular cancer. The overall risk of cancer is 85% in PJS [56,84].

PJS is an autosomal dominant condition with inactivating mutations in the *STK11* gene located on chromosome 19p13.3 (10 exons)[85]. Up to 80 % of cases are have mutations in the *STK11* gene including small insertions, deletions, splicing defects, nonsense and missense mutations and in around 30 % part or whole gene deletions are detected. The gene encodes a ubiquitously expressed multitasking serine–threonine kinase, which plays a critical role in several cell functions, including proliferation, cell cycle arrest, differentiation, energy metabolism, and cell polarity [56].

Juvenile Polyposis syndrome

Juvenile polyposis (JPS) is a heterogeneous, childhood to early adult-onset rare syndrome. JPS is characterized by the occurrence of juvenile polyps throughout the intestinal tract, mostly in the colorectum and patients have an increased risk of CRC. The diagnostic criteria for JPS are >5 juvenile polyps in the GI tract and/or any number of juvenile polyps with a family history of JPS. Polyps with adenomatous dysplasia might also be present. Lifetime risk of CRC has been estimated to be 40% to 70% [86,87]. Patients with Cowden syndrome can present multiple juvenile colonic polyps and therefore be misdiagnosed as having JPS [60].

JPS is an autosomal dominant condition caused by inactivating mutations including truncating mutations, splice site mutations and large abbreviations, mainly in two genes involved in the BMP/TGF-β signaling pathway, *BMPR1A* (chr10q22) and *SMAD4* (18q21) [88,89]. Around 15% to 60% of cases have mutations in the *SMAD4* gene and 25%-40% have mutations in the *BMPR1A* gene [90]. Mutations in the endoglin gene (*ENG*) have been found in a small proportion of cases around 2% [91]. The large variability in the mutation frequency reported likely reflects the small number of patients reported in each study. The *SMAD4* gene encodes a protein that is a mediator in the signaling from the TGF-β and BMP receptors on the cell surface to the nucleus. BMPR1A is a serine-threonine kinases type I receptor of the TGF-beta superfamily that when activated lead to phosphorylation of *SMAD4*. Mutations in *SMAD4* and *ENG* are also associated with hereditary hemorrhagic telangiectasia (HHT).

Cowden Syndrome

Cowden syndrome (CS) an autosomal dominant inherited syndrome is part of the phenotypically diverse spectrum of syndromes with germline mutations in the *PTEN* gene collectively called PTEN Hamartoma Tumor Syndromes (s) (PHTS). Hamartomatous gastrointestinal polyps occur throughout the gastrointestinal tract with the most frequent site being the stomach, colon, esophagus and duodenum. A mixture of polyps with different histology is common including adenoma, hamartoma lipoma, ganglionneuroma-line, juvenile and inflammatory polyps and the number can range from none to innumerable. Around 85% of CS patients have characteristic cutaneous facial lesions and other craniofacial abnormalities and extraintestinal manifestations are common. Soft tissue tumors include lipomas, hemangiomas and neuromas. The risk of CRC is 13% and the patients also have an increased risk of breast cancer (25%-50%), thyroid cancer and endometrial cancer (10%) [84,92].

CS is caused by germline mutations in the phosphatase and tensin homologue (PTEN) tumor supressor gene located on chromosome 10. It has multiple and yet incompletely understood roles in cellular regulation. The protein is known to signal down the PI3K/Akt pathway and cause cell cycle arrest and apoptosis and the protein has also been shown to regulate cell-survival pathway like the mitogen-activated kinase (MAPK) pathway. PTEN may play a role in cellular migration and focal adhesion, which then

include all processes that are important for normal cellular growth [93]. Inactivating mutation include small insertions, deletions, splicing defects, nonsense and missense mutations as well as part or whole gene deletions. Mosaic mutations have also been found in this syndrome [23,24].

Hereditary Mixed Polyposis Syndrome

Hereditary mixed polyposis syndrome (HMPS) is characterized by the presence of mixed polyps of several histotypes, but the main part resembles adenomas. Patients can also have juvenile like polyps and serrated adenomas in the colon and rectum, but absence of upper gastrointestinal abnormalities. The phenotype may overlap with JPS and might in some cases be indistinguishable. There is also an increased risk of CRC. All HMPS families reported so far is compatible with a dominant inheritance. The genetic defect in the first HMPS family described in 1997 was recently identified. It was found to be caused by a duplication of 40 kb upstream of the *GREM1* gene, which is an antagonist in the BMP signaling pathway [94]. However, families classified as having HMPS have also been shown to carry mutations in *BMPR1A*, these patients are presented with juvenile polyps, which were absent in the first HMPS family described. [95].

Serrated polyposis syndrome

Serrated polyposis syndrome (SPS) formally known as hyperplastic polyposis syndrome is a relative rare cancer syndrome characterized by multiple serrated polyps of the colon. SPS has been associated with an increased risk of CRC. Three categories have been distinguished: hyperplastic polyps, sessile serrated adenomas and traditional serrated adenomas. The genetic base of SPS is to a great extent unknown, but both dominant and recessive inheritance has been proposed, and there probably exist more than one genetic cause of SPS. Recently germline nonsense mutations were found in the *RNF43* gene in patients who presented with sessile serrated adenomas. The *RNF43* gene is a negative regulator of the Wnt signaling pathway [96].

Polymerase Proofreading Associated Polyposis

Polymerase Proofreading Associated Polyposis (PPAP) was recently identified as a new polyposis syndrome characterized by 10-100 adenoma with or without CRC (similar to MAP) or early onset CRC, some individuals also present large adenoma (similar to Lynch syndrome). Adenomas are the most common polyps, but hyperplastic polyps have also been found. Unlike tumors in Lynch syndrome most tumors with germ-line *POLE* or *POLD1* mutations are microsatellite stable. Patients seem to be at higher risk for development of other cancers, but since the number of reported families is still very low, the risk of CRC as well as other cancers has yet to be determined [97,98].

PPAP is an autosomal dominant condition caused by heterozygote mutations in the *POLE* and *POLD1* genes. The genes are large (*POLE* 49 exons, *POLD1* 27 exons) and

include a catalytic and an exonuclease domain. The number of mutations identified so far in *POLE* and *POLD1* is limited, but almost all of them occur in the proofreading (exonuclease) domain of the two proteins [97,99]. The first mutations identified were *POLE*, p.Leu424Val, and *POLD1* pSer478Asn by Palles et al [97]. Several SNPs are located in conserved sites within the catalytic or proofreading domain of *POLE* and *POLD1*, but no cancer risk has been associated with these in genome-wide association studies. A common polymorphism in *POLD3*, however have been found to associate with an increased risk of CRC in the northern European population.[100,101].

POLE comprise the catalytic and proofreading subunit of the leading strand DNA polymerase ε and *POLD1* comprise the catalytic and proofreading subunit of the lagging strand DNA polymerase δ. DNA polymerases are responsible for synthesis of DNA and are essential for replication. DNA replication involves multiple enzymes and pol δ and pol ε perform the bulk of the replication. The proofreading domains of *POLE* and *POLD1* enzymes ensure that these polymerases have very low error rate. Both of the enzymes are hetero tetramers and are ubiquitously expressed. They show a high level of evolutionary conservations and are homologous over their exonuclese domains. Apart from DNA replication they play essential roles in repair of chromosomal DNA involved in several pathways including nucleotide excision repair (NER), mismatch repair (MMR) and repair of double stranded breaks (DSBR) [100,102,103] (Figure 7).

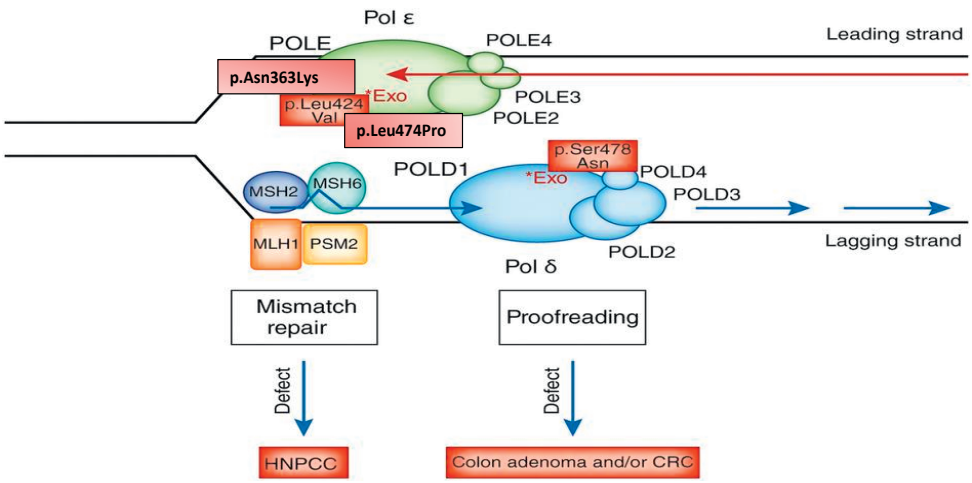

Figure 7. *Inherited DNA replication defects lead to CRC risk and PPAP syndrome and defective in mis-match repair protein lead to HNPCC and Lynch syndrome. Proofreading-impaired variants located in the exonulease (exo) domain of POLE or the POLD1 subunit of DNA polymerases identified by Palles et[97] (POLE, p.Leu424Val, and POLD1 pSer478Asn) as well as the two POLE mutations p.Asn363Lys and pLeu474Pro identified in paper IV and V respectively.(Reprinted and modified by permission from Macmillan Publishers Ltd: Nature Genetics] [103] © (2014).*

Lynch syndrome and Familial Colorectal Cancer type X

Lynch syndrome represent up to 50% of HNPCC (hereditary non-polyposis colorectal cancer) cases and account for 2% to 4% of all hereditary CRC syndromes. Individuals with Lynch syndrome are predisposed to various types of cancer especially colon and endometrial, but also ovarian, stomach, small intestine, hepatobiliary tract, upper urinary tract, brain and skin. Affected individuals can develop colonic adenomas with greater frequency than the general population but polyposis is rare. Specific criteria called the Amsterdam criteria and the Bethesda guidelines are used to identify families with Lynch syndrome. CRC arise at a younger age and at a more proximal location compared with sporadic CRC. Lifetime risk of CRC is estimated to be 50% to 80%. Endometrial cancer is the most common extra colonic malignancy associated with Lynch syndrome with a life time risk of 40%-60% [43,55,56,60]. Familial colorectal cancer type X (FCCX) families fulfils the Amsterdam criteria, but they have MSS tumors and no mutations are found in the MMR genes. This is a heterogenic group which seems to have a lower incidence of cancer and lower risk for non-colorectal cancers than families with documented DNA mismatch repair deficiency [104,105].

Missmatch repair genes and mutations

Lynch is an autosomal dominant syndrome and it is caused by mutations in the mis-match repair genes (MMR) genes, including *MSH2*, *MLH1*, *MSH6*, and *PMS2*. The MMR system is necessary for maintaining genomic fidelity by correcting single-base and insertion deletions mis-matches during replication. Mutations in *MLH1* and *MSH2* account for up to 70%, mutations in *MSH6* for 20%, and in *PMS2* for 10%. Most mutations found are truncating substitutions, a quite large proportion, approximately a third, are of missense type and often require functional tests for the assessment of pathogenicity. A five-tiered system to classify MMR gene variants for pathogenicity was recently incorporated in the InSiGHT database [12].

Constitutional *MLH1* promoter hypermethylation is another mutation type found in Lynch syndrome. Methylation of a single allele in all tissues resulting in complete silencing of the gene is often the case, but mosaic methylation in blood and partial silencing can also occur [106-108].

In 2006 germ-line deletions of the 3´part of the epithelial cell-adhesion molecule gene, *EPCAM,* were found in a subset of families with Lynch syndrome. The deletions were subsequently found leading to transcription of *EPCAM* lacking a stop codon and resulting in a fusion transcript between *EPCAM* and *MSH2*. The consequence was transcription into the adjacent structurally normal *MSH2* gene and initiation of hypermethylation of the *MSH2* promoter with the subsequent silencing of the gene. The *MSH2* promoter is only methylated in tissues expressing *EPCAM*, which mainly are epithelial cells [109,110].

Microsatellite instability testing

There are several laboratory based strategies that help establish the diagnosis of Lynch syndrome including testing the tumor tissue for the presence of microsatellite instability (MSI) and loss of protein expression for any one of the MMR proteins by immunohistochemistry (IHC). MSI phenotype is however not restricted to inherited cancer cases, around 15%-20% of sporadic colon cancers are MSI. MSI testing can therefore not fully distinguish between a somatic (sporadic) and a germline (inherited) mutation. Defective MMR in sporadic cancer are most often due to *MLH1* promoter inactivation by hypermethylation. A specific mutation in the *BRAF* gene, V600E (Val600Glu) has been shown to be present in 70% of tumors with hypermethylation of the *MLH1* promoter, whereas the V600E mutation is rarely identified in cases with germline *MLH1* mutations. Assessment of *MLH1* promoter methylation and testing for *BRAF* V600E mutation can therefore be used to help distinguish between a germline mutation and epigenetic/somatic inactivation of *MLH1*. Tumors that have *BRAF* V600E mutation and *MLH1* promoter methylation are almost certainly sporadic, whereas tumors that show neither are most often caused by an inherited mutation [111,112].

MMR proteins

The components of the mismatch repair system (MMR) are highly conserved in both pro- and eukaryotic organisms. MMR proteins can recognize both single nucleotide mis-matches and mis-matches caused by insertion and deletions. In humans single base mismatches are recognized by a heterodimeric complex of MutS related proteins: MSH2-MSH6. Insertion and deletion of 2-8 bases can only be recognized by MSH2-MSH3. There is an overlap in the specificities of these two complexes and some redundancy in there activity. Mis-match binding is followed by assembly of MutL related proteins MLH1-PMS2 and another alternative complex formed by MLH1-MLH3, which bind to the MSH containing complexes together with replication factors and other proteins to proceed with excision and resynthesize. MSH2 and MLH1 are the common components of these complexes and inactivation of either will abolish the MMR activity whereas loss of one of the other components will only diminish the activity. MMR components also interact with proteins involved in nucleotide excision repair (NER). POLD 1 is also thought to play a role in new strand synthesis as part of NER and MMR [43,112].

Moderate and low penetrant loci and variants

Family history is a major risk factor for CRC, however, germline mutations in high penetrant genes account so far only for a small part (~5%). The remaining inheritance might be caused by a large number of low-penetrant loci that are common in the population. Studies have identified ten loci that confer a modest risk of CRC [113]. The ten tagging SNPs are located in intergenic and intronic areas that may affect gene expression through distant regulatory element. Several of these loci are located close to

members of the TGF-β superfamily signaling genes including, *SMAD7, GREM1* [114], *BMP2, BMP4, RHPN2*. Even though these ten variants independently only confer a low risk their additive contribution can be much higher [115]. There also exist a number of variants that have been associated with an increased CRC risk in known high risk genes like the APC-Ile1307Lys, which in the Ashkenazi Jewish carriers have been estimated to confer an 2-fold increased CRC risk in carriers [116]. Increased risk has also been found for variants in the *TCFL7, GALNT12* and *TGFBR1* gene among others [117-119].

Other high penetrant genes

When it comes to high-risk families like FCCX , the gene or genes are likely uncommon but sufficient penetrance to give rise to the observed autosomal-dominant segregation patterns. The genetic loci identified through familial linkage analysis do not overlap with the susceptibility loci identified through genome-wide studies. Recently a truncating variant in *RPS20* was identified in an FCCX family, segregating with affected individuals, however, this mutation or any other mutations in this gene have so far not been found in other families[120]. Another study of FCCX kindred identified germline variants in the *SEMA4A* gene predisposing to colorectal cancer [121]. There could also be other types of inactivation mechanisms in these families possibly involving e.g. epigenetic modifications.

OBJECTIVES

The overall aim of this thesis was to identify and characterize mutations in families with hereditary colorectal cancer by using different methods. Paper I and II focus on mutation detection in the *APC* gene, mainly in families with a classical FAP phenotype or AFAP. Paper III reveals the sensitivity of five screening methods used including MPS for mosaic mutation detection and paper IV-VI focus on identification of new predisposing genes with targeted MPS (whole exome and panel-based sequencing).

Specific aims

Paper I
Mutation screening of the *APC* gene and clinical characterization of 96 unrelated FAP patients from the Swedish Polyposis Registry, by use of a variety of different methods and performing genotype to phenotype correlations.

Paper II
To identify the causative mutation in the largest FAP family (Family 1) in the Swedish polyposis registry and determine the significance of promoter 1B in the *APC* gene.

Paper III
Compare the sensitivity for mosaic mutation detection by using clinical diagnostic screening methods including massively parallel sequencing (MPS).

Paper IV
To identify the causative mutation in three AFAP families without mutations in known colorectal cancer predisposing genes by use of whole exome sequencing (WES). Identified genes with nonsense or frameshift mutations and also selected missense mutations in known CRC-associated predisposing genes were also investigated in 107 patients divided into 11 clinical subgroups. The aim of this study was to identify high or moderate penetrant variants in these patient subgroups.

Paper V
To identify the causative mutation in a large colorectal cancer family with a multi-tumor spectrum.

Paper VI
To identify the causative mutations and classify missense mutations in 76 clinical sub grouped colorectal cancer patients by use of a targeted panel including 19 CRC-predisposing genes.

MATERIAL AND METHODS

Material

In this thesis DNA from blood lymphocytes from patient and control have been used in all papers I-VI. In addition, in paper II tissue from polys and normal colon mucosa were used from four individuals and a panel of different tissues (total RNA Master Panel II (Clontech Laboratories). In paper IV normal colon mucosa was analyzed from two individuals in Family C, controls included normal colon mucosa from individuals with sporadic colorectal cancer and one purchased normal colon mucosa control (Ambion). In paper V tumor DNA was analyzed from two individuals.

Basic methods

Polymerase chain reaction (PCR)

PCR was introduced in the mid-1980s[122] and is the main molecular technique used to amplify a defined target DNA sequence. To permit this selective amplification two oligonucleotide primers (15-25 nucleotides) one forward and one reverse, complementary to the specific target sequence are needed. The design of primers is often done by using different software. In this thesis DNASTAR from Lasergene or UCSC primer design was used.

A PCR cycle consists of three steps; 1) denaturation of the ds DNA template, 2) annealing of primers to single stranded template 3) DNA synthesis (elongation), which is a chain reaction, were newly synthesized DNA will act as a template in the next cycle. The DNA is amplified exponentially in the presence of primers, deoxyribonucleotide triphosphate nucleotides (dNTPs), buffer and heat stable DNA polymerse whereby millions of target DNA are produced, Figure 8 a.

The PCR run can be divided into three phases (exponential, linear and plateau), where in traditional PCR the product of the final phase is measured, this is called end-point PCR. In real-time PCR, the amount of PCR product is monitored in each cycle and the data is measured during the exponential phase of the PCR reaction (see expression methods). The length and amount of PCR product traditionally is measured on an agaros gel, but dye labeled primers (Fam, Tet etc) can be used as well, which make detection on an automatic capillary electophoresis instrument necessary.

In traditional PCR (Figure 8a) a sample offers a single measurement, where only average signals from the two alleles can be calculated and an allele frequency can only be estimated from the ratio between allele one and allele two. In a digital PCR every template is being amplified separately and the amount of allele one and two can be

calculated precisely making an absolute quantification possible. This clonal amplification is the basic method in massively parallel sequencing [123] Figure 8b.

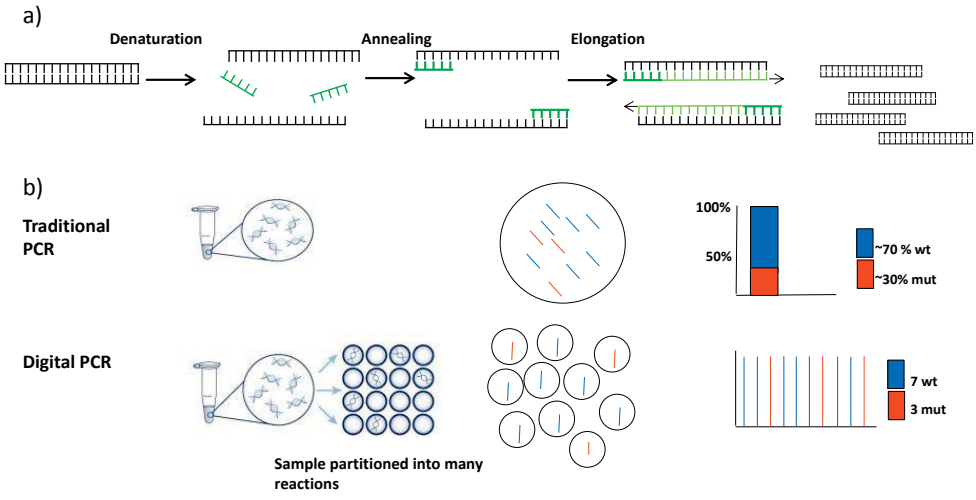

Figure 8. *a) PCR amplification steps b) Traditional PCR and digital PCR*

Previously used methods

Basic methods previously used in genetic testing are single strand conformation polymorphism or Heteroduplex (SSCP/HD), Denaturing High Performance Liquid Chromatography (DHPLC) and Protein truncation test (PTT). All of these were used in paper I and III and served their purpose well at that time.

Sequencing methods

From Sanger sequencing to massively parallel sequencing (MPS)

Sequencing has been going through a revolution in the last years. Historically the most common method of sequencing, originally developed by Sanger et al in 1977[124], is called the Sanger method or dideoxy method. A further development of this is cycle sequencing in which the template for the reaction is a purified PCR product. One primer is added in each reaction, either forward or reverse, together with deoxynucleotides (dNTPs), DNA polymerase and dideoxynucleotides (ddNTPs). DdNTPs have exchanged the 3´OH group needed for chain elongation with a hydrogen atom and are fluorescently labeled. In each reaction ddNTPs are randomly incorporated and terminates the elongating chain, this mixture of fluorescently labeled DNA strands are after purification separated by automated capillary electrophoresis according to size. The fluorescently labeled DNA allows for detection and basecalling by software programs. By using this sequencing method a mixure of DNA molecules

are sequenced. In each position the base with the strongest signal is called. In this thesis the sequencing chemistry supplied by Life Thechnologies (BigDye) was used after purification of the PCR products with or ExoSAP-IT (USB) or AMPure magnetic beads (Agencourt Bioscience Corporation). Sequencing was performed on the Genetic Analyzers 3100, 3130XL, 3730 (Life Technologies) in house or at GATC Biotech AG, European Custom Sequencing Center (Germany). All sequencing in paper I and II was performed with Sanger and confirmation of the results from massively parallel sequencing in paper IV, V and VI.

General principles of Massively Parallel Sequencing (MPS)

Beginning in 2005 the traditionally Sanger-based approach to DNA sequencing has gone through revolutionary changes and a new era with Massively Parallel Sequencing (MPS) more commonly named Next Generation Sequencing (NGS) has been taking over. One important difference compared to Sanger sequencing is that MPS is a digital sequencing method and each sequence named read originates from a single DNA molecule. For targeted re-sequencing there are mainly two technical preparation methods to construct a library of fragments. The first is by PCR involving multiple primer pairs in a mixture that are combined with genomic DNA in a multiplex PCR. Primers include all sequence needed for downstream analysis (see adapters). The second approach is a two-step process, first a sample preparation step of the genomic DNA and then a fishing step were the genomic regions of interest are extracted by a hybridization-based technique. The library construction in the second approach starts with DNA fragmentation, that can be done either by restriction enzym digests, by using transposase or by using the covaris which is a focused ultrasonicator. Then synthetic DNA adapters are covalently added to each fragment end by DNA ligase. The adapters in general contain three parts: universal sequences specific to each platform, sequences used to amplify the DNA and indexes (also called tags or barcodes) specific for each sample. These are used for distinguishing samples when pooling several samples together, which can be done before capture or after capture. There are in general two main types of indexes, in-line or multiplex, were in-line indexes are adjacent to the samples DNA and read from the same sequencing primers as part of the reads. The multiplex barcodes are located in the stem of the library between the two universal sequences.

The libraries are amplified (isothermal) *in situ* on a solid surface either on beads or a flat glass of micro fluidic channels with covalently attached adapters with sequences that are complementary to those on the library fragments. The amplification is digital, which means that each DNA molecule is amplified in a cluster or on a single bead that is delimited from other DNA molecules. This amplification is necessary to provide sufficient signal from each fragment. Sequencing from both end are achieved through pair-end sequencing (see Illumina sequencing for explanation). In MPS, sequencing and detection is done simultaneously and include three steps: addition of a nucleotide, detection of the nucleotide incorporated and finally a washing step that may include

chemistry to remove fluorescent labels or blocking groups. The reads length obtained from MPS in general 50-300 bp depending of the chemistry and platform used.

Emulsion PCR and pyrosequencing (454/ Roche)

The 454 was the first sequencing system for MPS commercially introduced in 2004 and use a sequencing technology known as pyrosequencing (paper III). It starts with a common library preparation step from genomic DNA or PCR products, as described above. In paper III separately amplified PCR products were used and the specific adapter sequences were included in the first amplification. Then an emulsion PCR takes place were the library fragments are mixed with a population of agarose beads covered with fragments of sequences complementary to the adapters and each bead is associated with a single fragment. Fragment bead complexes are then isolated into individual (oil:water) micells that also contain PCR reactants. Thermal cycling (emulsion PCR) produces around 1 million copies of each DNA fragment on the surface of each bead (Figure 9). The beads are put into a picotiter plate that holds a single bead in each of several hundred thousand single wells, which provide a fixed location at which each sequencing reaction can be monitored. Enzyme containing beads that catalyze the pyrosequencing reaction step are then added and the mixture is centrifugated to surround the agaros beads. The picotiter plate act as a flow cell in the instrument, in which one nucleotide at the time is introduced stepwise (Figure 9). Each incorporation of a nucleotide by the DNA polymerase results in the release of a pyrophosphate which is used in a coupled enzyme reaction in which the firefly enzyme luciferase will produce light. The emitted light is recorded by an imaging step with a CCD camera after each nucleotide incorporation step. The first four nucleotides TCGA on the adapter fragment adjacent to the sequencing primer is necessary for the base-calling software to calibrate the light emitted by single nucleotide incorporation[125] [126].

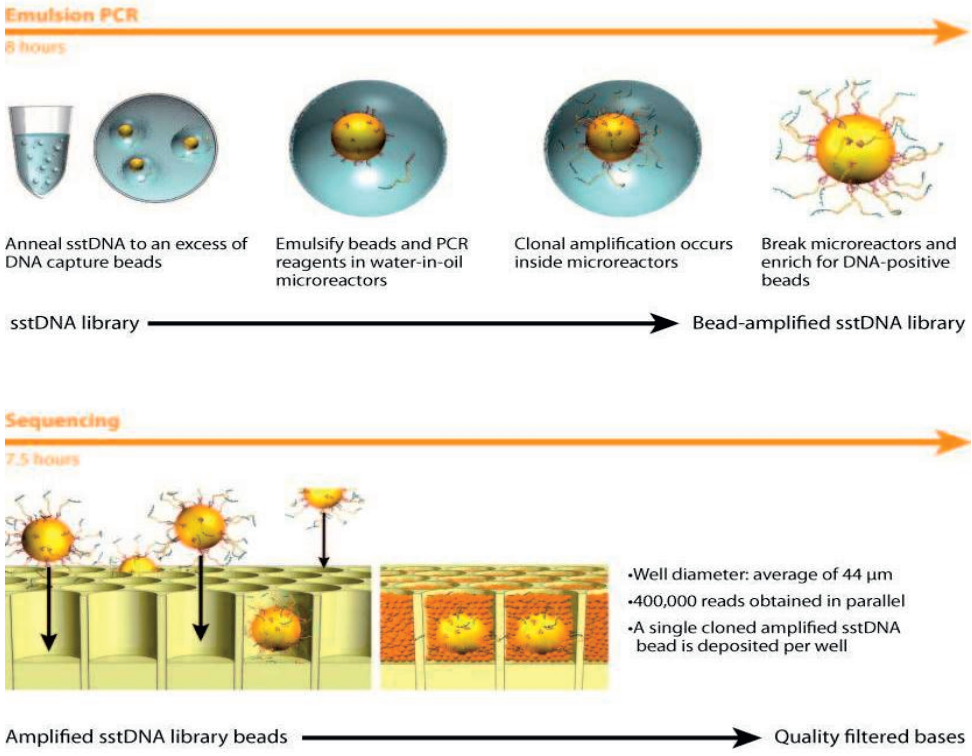

Figure 9. *The Roche/454 amplification and sequencing method (Reprinted from [127])*

Library preparation based on hybridization

The SureSelect library preparation kit (Agilent) was used in papers IV, V and VI, though in paper VI a modified version with in-line indexes was used. The library preparation starts with a DNA sample preparation step followed by a hybridization step. Fragmentation by covaris produce fragments with a base pair peak of around 200-300 bp, followed by end repair, phosphorylation and addition of A overhang. In the hybridization step probes (RNA baits) 120 nt long are used to extract either the whole exome (paper IV and V) or in the custom kit the 19 gene regions of interest (paper VI). A 16-24 hours hybridization is conducted followed by washing steps and dilution to prepare the samples for sequencing (Figure 10).

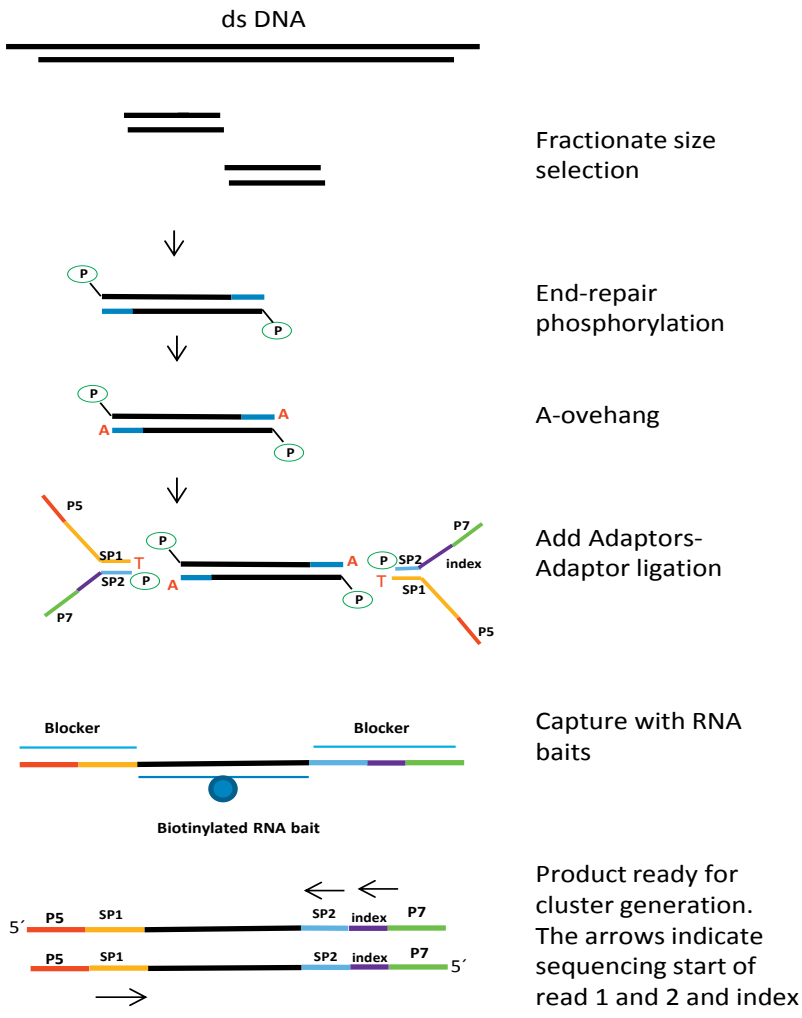

Figure 10. *Schematic overview of library preparation and capture process (According to the Sureselect Agilent).*

Bridge amplification and sequencing by synthesis (Illumina)

This sequencing system was initially developed by Solexa in 2007 and subsequently acquired by Illumina. The Illumina flow cell is composed of flat glass with eight microfluidic channels with covalently attached adapter sequences complementary to the library adapters. By careful quantitation of the library concentration a very precisely diluted solution of library fragments is amplified *in situ* on the surface of the flow cell by using bridge amplification to produce clusters of clonal sequences. First the fragments are denaturated, followed by hybridization to the two different oligos of the flow cell. These oligos work as primers and are strand complementary to the adapter sequences (T5 and T7) and an initial extension occurs. After another denaturation the first annealing cycle, where the fragments bend over to form a hybridized bridge to a nearby adapter, takes place. This goes on until enough molecules in each cluster have been generated. Finally, linearization of the fragments occur where the ends carrying the same adapters are released (Figure 11).

The reversible dye terminator sequencing starts with priming of the fragment with a complementary DNA primer. All four nucleotides, carrying fluorescent labels, are added in each cycle. The sequencing by synthesis occurs by the addition of single nucleotides containing a block at the 3´-OH position of the ribose suger, preventing additional base incorporation reactions by the polymerase. After a nucleotide has been added by the DNA polymerase, unincorporated nucleotides are washed away, the fluorescent groups are chemically cleaved and the 3´OH group is deblocked. The cycling reaction is repeated for up to 150 bp depending on the chemistry used. When performing paired end sequencing reading from the opposite end of each fragment then begins. This starts with the removal of the synthesized strands by denaturation and then regeneration of the cluster by performing limited bridge amplification. After this amplification step the opposite end of the fragment is released and the fragments are primed with reverse primers and sequencing is conducted as described [125]. Paired end sequencing was used in paper IV, V and VI.

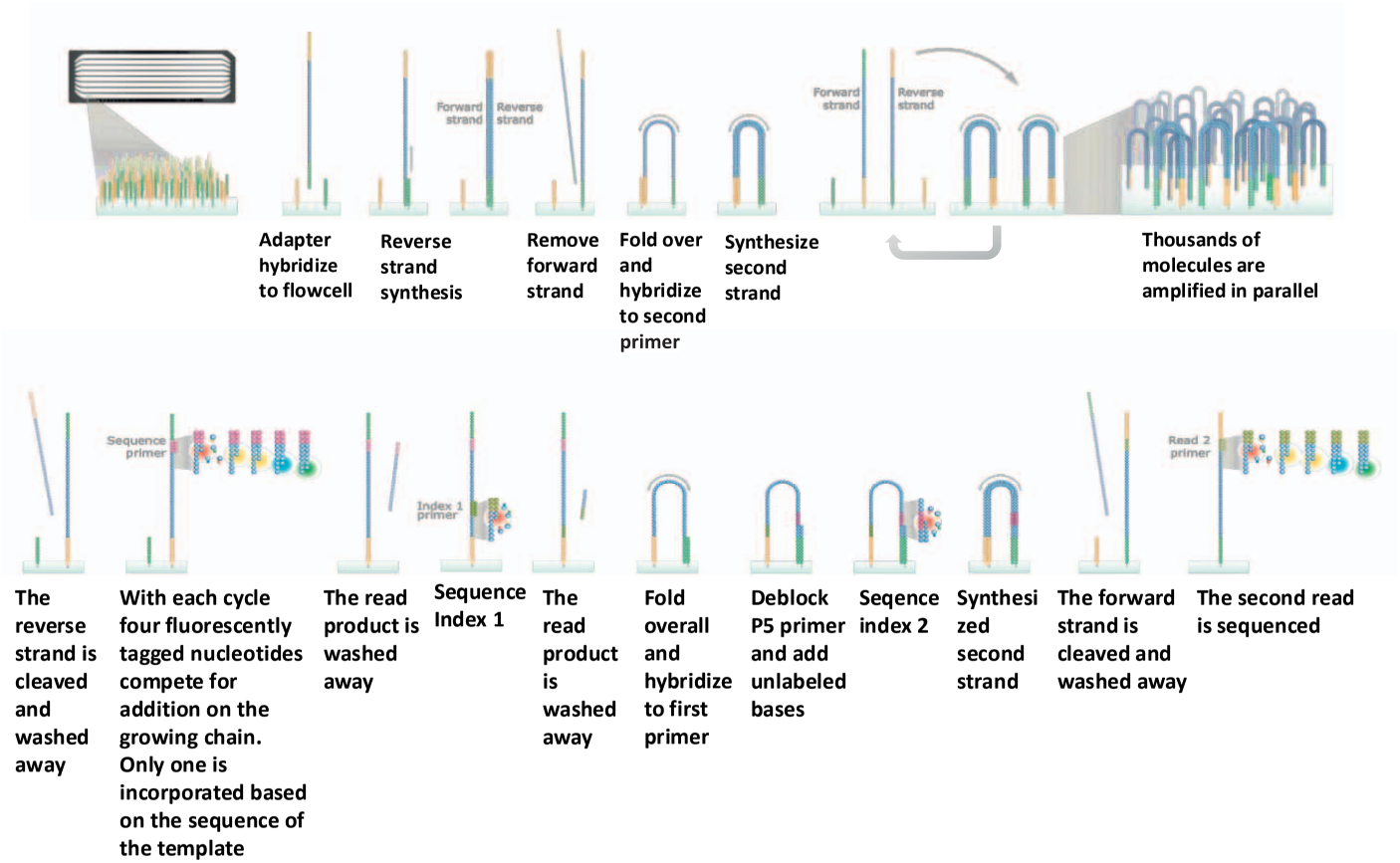

Figure 11. Bridge amplification and sequencing by synthesis (from Illumina Inc ©2014)

Limitations by noise; Advantages/Disadvantages 454 and Illumina sequencing

Depending on the platform the contribution to noise in the sequencing reaction differ and there is an interplay between the sources of noise and the sequencing errors that may result; this is instrument and chemistry specific where both read length and error types can be considered. Errors can also be independent of instrument as the use of PCR itself can contribute to systematic errors during the library construction. This was indeed seen in paper III, however this can be influenced by using a high-fidelity polymerase and/or by limiting the number of amplification cycles independent of the instrument used.

The pyrosequencing technology (454) allows for sequencing read length up to 700 bp with accuracy 99.9% [128]. A great disadvantage however is that the basecalling cannot properly interpret long stretches of the same nucleotide, homopolymers, which make these regions prone to base insertions and deletions errors. In contrast because each incorporation step is nucleotide specific, substitution errors are rarely encountered in 454 sequencing reads. Illumina does not appear to share these limitations but it has its own systematic base calling biases. Different tiles of the sequencing plate tend to produce reads of different quality [129] and the 3' ends of sequences tend to have higher sequencing error rates as compared to the 5' ends [130]. Decreased accuracy with increasing nucleotide addition steps and increased single-base errors (substitution errors) have been observed in association with GGC motifs [131]. Sources of noice include phasing where increasing number of fragment fall out of phase with the majority of the fragments in the cluster due to incomplete deblocking in prior cycles or due to lack of blocking groups that allow an additional base to be incorporated. Additional sources are residual fluorescence interference noise due to incomplete fluorescent label cleavage from previous cycles.

The power of coverage

In MPS the sequencing cost sets the limit to the amount of sequences that can be generated and consequently the biological outcome that can be achieved from an experimental design. The term coverage and reads depth is used in these calculations. The theoretical or expected coverage is the average number of times that each nucleotide is expected to be sequenced given a certain number of reads of a given length if the reads are randomly distributed across a target region The coverage can be calculated by C=LxN/G were C= coverage, L= read lengh, N= number of reads and G is the length of the target region. The depth refers to the average number of times that a nucleotide is represented in a collection of random raw sequences. Coverage and depth can be used interchangeably. The coverage is important for the variant detection, which is reduced by low base quality and by non-uniformity of coverage. Increased sequencing depth can improve these issues and thereby reduce the false-discovery rate for variant calling. Coverage can be affected by sample preparation. The main sources are GC biases that are introduced during DNA amplification. Uniformity of coverage will also be influenced by low-complexity sequences, which restricts bait design or lead

to off target capture [132]. Coverage is also used when dealing with the breadth of coverage of a target region (horizontal coverage). This is defined as the percentages of target bases that are sequenced a given number of times. For exome studies a minimum of 80% of the target to be covered by at least tenfold have been used and a mean on target read depth of 20x. [133].

Bioinformatics

Data processing

The processing of the output data from MPS compromise several step and can be done by use of many different algorithms, but include the main steps: basecalling, quality control, aligment and variants calling (Figure 12). Basecalling is generally done on the instrument. By use of noise estimating, a measurement of uncertainty to each base call from image analysis, base quality scores are produced by the base-calling algoritms in addition to identifying nucleotides. These quality values are usually given in the standard Phred quality score (Phred score 20 corresponds to a 1% error rate in base calling). Demultiplexing of samples can often be done on the instrument as well if multiplex indexes are used, in-line indexes need to be handled by post-image processing.

Checking the quality of the generated read data is often the first step in a pipeline. This can be done by various programs, for example FASTQC. Checking for over-represented sequences, deviation from expected GC content, distribution of nucleotides per read position allows for fast identification of problems that can occur during sample preparation and sequencing. Read trimming can often be done by removing bases at the end of the read that are likely to contain sequencing errors, increasing the number of mappable reads.

Alignment

Alignment of reads can be done through *de novo* assembly or to a reference genome. In paper IV, V and VI read alignments were done to the human reference genome GRCh37/hg19 or hs37d5ss (1000 genome with decoy sequences). The aligner used in paper IV and V was BWA, which is a short read aligner based on an algorithm called Burrow-Wheeler transform, this is a fast memory efficient aligner [134]. BWA is commonly used both in bioinformatics pipelines and in software programs. In paper VI Novoalign was used as this at the time was found to produce the most accurate overall results.

The accuracy of the alignment has an important role for the variant detection. The aligner has to produce well calibrated mapping quality scores as the probability of an observed variant call depend on those scores. It is further important also for the aligner to be able to deal with sequencing errors as well as real differences between the reference genome and the sequenced genome. The amount of sequence identity

required between each read and reference sequence is determined by a trade-off between accuracy and depth.

Once the reads have been aligned to the reference genome the results are often stored in the sequence alignment/map (SAM) format. The SAM format store information about each aligned read including the position on the reference contig, the orientation of the read, quality of the alignment and potential further alignment possibilities of the read. The binary version of SAM is called BAM and the SAM/BAM format has become the standard format for storing the results of the alignment steps and can be used by down-stream toolkits. Visualization of the alignment can be done in the Integrative Genomics Viewer (IGV). Before proceeding with the variant calling post alignment, processing is usually done. The PCR used for amplifying the library may introduce artifacts, read-pairs mapping to identical genomic coordinates, likely to represent PCR copies of the same DNA template are commonly marked (duplicate marking) or removed (Piccard tools). Local realignment around small indels are usually done, differences in resolving indels may result in artificial SNPs in the downstream analysis, base quality recalibration is also done in which not only the raw quality score, but also the position of the base in the read are taken into account. Obtaining well-calibrated quality scores is important since SNP and genotype calling at a specific position in the genome depends on both the base calls and the per-base quality scores of the reads overlapping the position. Both the local realignment and the base quality recalibration in paper IV-VI have been done with the Genome Analyzer Tool Kit (GATK)[135].

Variant discovery and genotyping

SNP calling determine in which positions at least one base differ from the reference sequence and genotyping is the process of determining the genotype of each locus and this is only done for positions were variants already have been called. The accuracy for variant calling is highly dependent on mapping quality, read depth and allele balance. SNP calling and genotyping can be done simply by counting alleles at each site using cut-off rules for when to call a SNP, for example a heterozygote genotype is called if the proportion of the non-reference allele is between 20 % and 80 % otherwise a homozygote genotype would be called. Commercially available softwares use similar methods, however, several probabilistic methods have recently been developed. Probabilistic methods produce robust estimates of the probabilities of each of the possible genotypes, taking into account noise, as well as other available prior information that can be used to improve the estimation of the genotype. The most used are also provided by the GATK, in which currently two variant callers are available, the UnifiedGenotyper and the HaplotypeCaller. In paper IV-VI GATK the UnifiedGenotype has been used which uses a probabilistic (Bayesian likelihood) model to estimate the most likely genotype incorporating the statistical uncertainty. [135,136]. The standard variant output format is called Variant Caller Format (VCF). The VCF file includes at least information about the chromosomal position, the reference base and the alternative bases for identified variants and the quality and depth. Filtration of

variants can be done based on the quality score, depth and strand biases to minimize artifacts.

Structural variations detection from MPS data

There are four general types of strategies to detect structural variants. These include read-pair methods, read-depth methods, split read approach and sequence assembly. Read-pair methods include the span and orientation of paired-end reads in which the mapping span and/or orientation of read pairs are inconsistent with the reference genome. Read pairs that map too far apart are defined as deletions and read -pairs that map too close are defined as insertions. Orientation of the reads can indicate inversion and tandem duplications. Read depth methods assume a random distribution of mapping depth and deviations from this are an indication of a duplication, where a higher read depth is seen, or a deletion for which a lower read depth is seen. Split-read approach define the break point of structural variants based on split sequence reads in which an alignment to the genome is broken- A continuous stretch of gaps in the read indicate a deletion or in the reference indicate an insertion. Sequence assembly is used to generate sequencing contigs which are compared to the reference genome [16,137,138].

In paper VI the detection of structural variations was based on read depth (coverage ratios). The span and orientation of the reads were evaluated in Integrated Genome Viewer (IGV). There is a relationship between the GC content of the pair-read fragments and the coverage of the targeted region and GC normalization is usually done to correct for this bias [139].

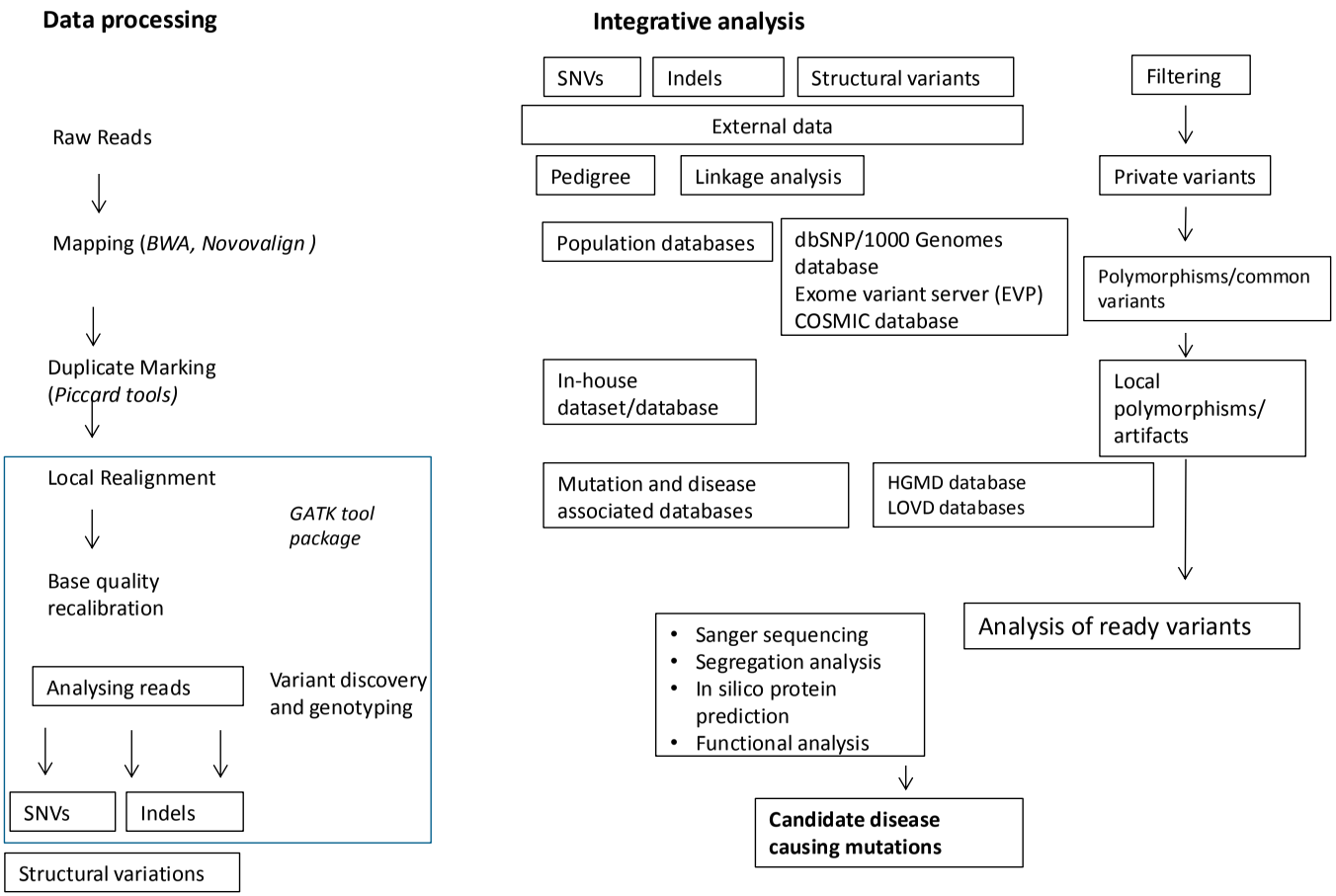

Figure 12 Data processing and integrative analysis

Annotation and Integrative analysis

Functional annotation is the process of identifying the locations/segments of genes , coding regions and other specific locations that are of importance in a DNA sequence and associated relevant information with those locations/segments in order to make sense of it. In paper IV-VI ANNOVAR was used for annotation[140]. To be able to filter for relevant data external data is often used. This includes pedigrees, which will give information about the disease model, sequencing of healthy individuals from the family will give information about private variants common in the family, if linkage data is available information regarding regions with highest LOD scores can be included to minimize the regions with potential disease-causing variants. Information from population databases provide information regarding minor allele frequencies and include db SNP and 1000 genomes, National Heart and Lung and Blood Institute (NHLBI) Exome Sequencing Project (ESP). COSMIC database among others can provide information about the presence of specific variants in tumors.

Local datasets or databases will both give information about local variants and in addition provide a filtering step for platform specific artifact. Confirmation of variants by another method (e.g. Sanger sequencing) and segregation analysis in the families are necessary for prediction of a pathogenic variant (Figure 12).

CNV methods

Multiplex Ligation-dependent Probe Amplification (MLPA)

MLPA is a semi-quantitative multiplex probe hybridization based technique develop by Schouten et al in 2002 [141]. It is a method for detection of deletion or duplication of exons or whole gene regions in DNA or mRNA. It can also be used for detection of methylated regions. In each assay probes for specific genes or chromosomal regions are included together with several additional control probes .The MLPA assay includes the following steps: 1) denaturation of DNA, 2) hybridization of two probes next to each other on the DNA template, 3) Ligation of the two probes 4) PCR amplification with universal primers forward and reverse, were the forward primer is fluorescent labeled, 5) amplified products are separate on capillary electrophoresis.

In paper II we used methylation-specific mlpa (MS-mlpa), which work in a similar way to standard MLPA (see section CNV methods) except that the target sequence contains restriction-site for the methylation sensitive endonuclease HhaI. Unmethylated samples will be digested when incubated with HhaI after hybridization and no amplification of these sample probe hybrids will occur, leading to no signal in the analysis. In contrast DNA methylated samples are protected against HhaI digestion amplification will occur and a peak will be present.

Data analysis

The areas of the resulting peaks are normalized by dividing each peak area by the combined peak areas of all peak areas in the sample analysis (specific and control peaks). Copy number of each target is determined by the comparison to a control samples after normalization. The data analysis in this thesis has been done in the program SeqPilot (JSI Medical System).

CNV detection based on read depth from MPS data

The CNV analysis in paper VI is based on read depth by using a software (Törngren et al. In prep) developed by Anders Kvist and Staffan Living (Department of Oncology. Lund). In this program one read-pair represents one data point in a sliding window over the target region. A normalized coverage depth ratio between a sample and an average of 23 normal samples (baseline) are computed, including GC-normalization. Detection of abnormal coverage ratios are found by visual inspection of plots of the coverage ratios over the targeted regions. Red dots are below or above the normal variation (cutoff 0.75 and 1.25). In regions with an abnormal coverage ratio a loss or a gain can be suspected (Figure 13). A manual comparison between samples for a specific region was done and the integrated genomics viewer (IGV) was used to define the breakpoints (Figure 14). In paper IV control free copy number caller (FREEC) was used. This method also uses GC content normalization to predict copy number alterations.

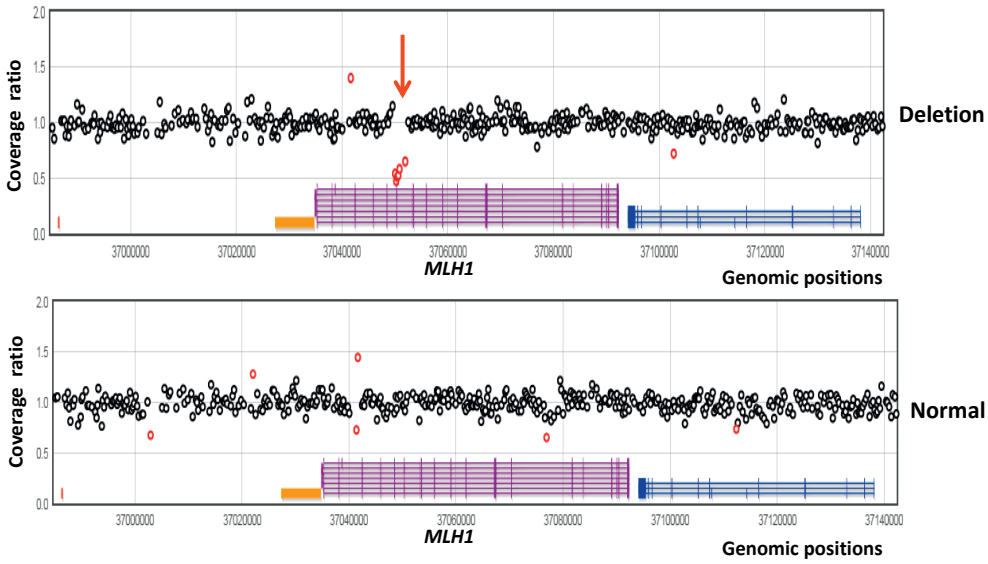

Figure 13. *The coverage software presenting a likely deletion as show in the upper window compared with the lower window of a sample with no abbreviations.*

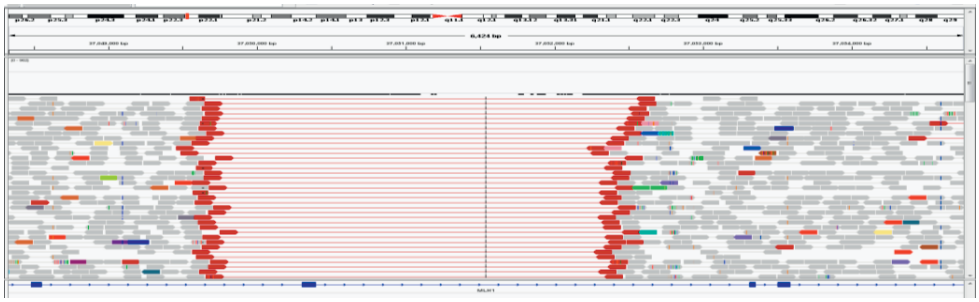

***Figure 14**. Part of the MLH1 gene region defining the breakpoints of the deleted region.*

SNP array analysis

SNP arrays are high density oligonucleotide-based arrays. SNP array probes comprise 25-mer oligonucleotides. They enable genotyping of SNPs as well as copy number changes including losses (deletions) and gains (duplications and amplifications), loss of heterozygosity (LOH) and copy neutral LOH of chromosomes or chromosomal regions. Separate probes are synthesized to match each of the possible alleles, enabling genotyping of the SNP by comparing fluorescent intensity between the two sets of probes. The probe intensities, that correspond to the two possible alleles of the SNP reveal which of the three genotypes possible (AA, AB, BB) that is present. These probe intensities can also be used to estimate copy numbers and in some arrays also non-polymorphic copy number probes are present. Hybridization is performed on a single sample per microarray. The fluorescent intensities are compared in silico to a set of reference samples from healthy individuals.

In this thesis Affymetix Genome-Wide SNP array 6.0 or cytoscan HD array were used in papers II, IV and V. The Genome-Wide SNP array 6.0 features 906,600 SNP probes and around 946,000 copy number probes, the average space between the probes are 1.6 kb and the cytoscan HD arrays have 750,000 SNP probes and 1.9 million copy number probes.

Total genomis DNA is digested with Nsp 1 and Sty 1 restriction enzymes and are then ligated to adapters separately. The products are pooled after a PCR amplification performed with generic primers. The PCR products are fragmented (DNasel) and end-labeled with proprietary biotin-labeled reagent by using the enzyme Terminal Deoxynucleotidyl Transferase (TdT) before hybridized to the array (Figure 15).

Data analysis

The data analysis in paper III, IV and V was done with softwares provided by Affymetrix (Genotyping Console v2.1 and Chromosome Analysis Suite 1.0.1 (ChAS)), where Hapmap controls provided by the software as well as in-house-controls has been

used for comparison and exclusion of common CNVs for the CNV analysis In paper IV and V genotyping was also performed for further linkage analysis.

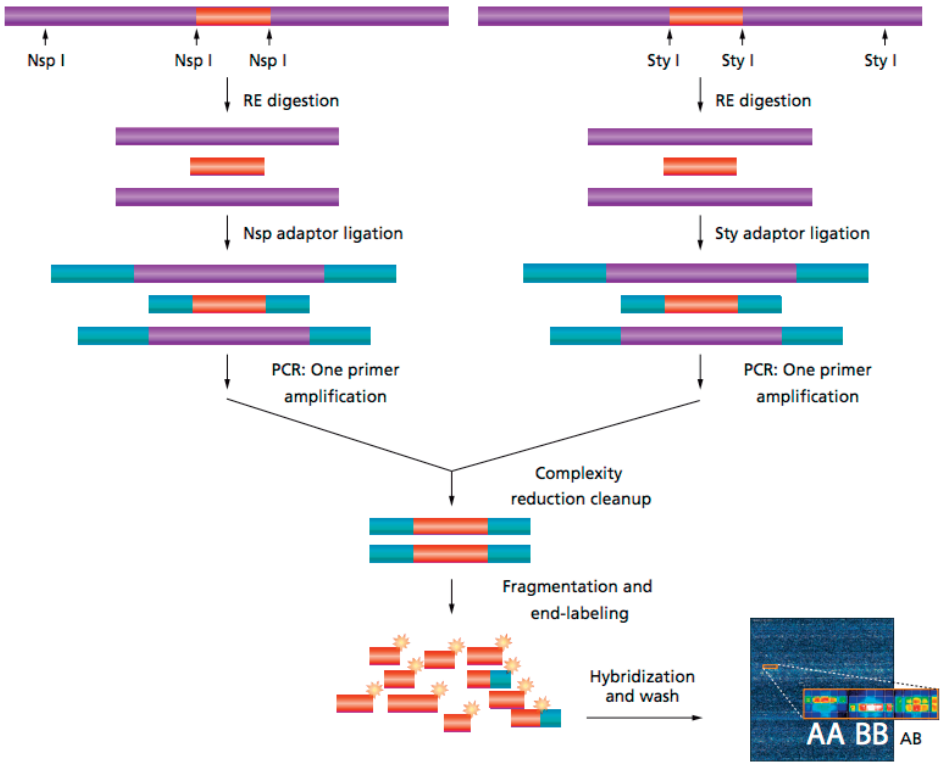

Figure 15. *Schematic overview SNP array workflow (Reprinted Affymetrix, Inc ©2014)*

Expression analysis methods

Real-time RT-PCR (Real-time Reverse Transcriptase PCR)

The amount of mRNA transcripts can be measured by RT-PCR. RNA is first reverse transcribed into complementary DNA (cDNA) using reverse transcriptase and often random primers. In the following real-time PCR, the PCR is monitored as the amplification process proceeds and the cDNA level is measured during the exponential phase of the PCR reaction either by using fluorescent probes (TaqMan probes) or DNA binding dyes (SYBR Green). The TaqMan probe specific for the target is labeled both with a fluorescence- tag reporter and a quencher. During the polymerase chain reaction process, bound probe is degraded by the 5´-exonuclease activity of Taq

polymerase releasing the fluorescent reporter from the quencher (Figure 16). The fluorescence release is directly proportional to the amount of PCR product generated in each PCR cycle. A threshold is set in the exponential phase of the amplification curve and the threshold cycle number (Ct) is read, this will give the highest precision of the cDNA amount. A low CT-value indicates high number of transcripts while high CT-value corresponds to low number of transcripts.

Quantification of the results obtained by TaqMan analysis can be analyzed with the standard curve method or the comparative CT method. By using the standard curve method an interpolation of the sample against a standard curve generated from a serial dilution of a standard sample (calibrator sample). To correct for differences in starting amount of RNA or differences in cDNA conversion a normalization has to be done with a reference gene (endogenous control). Housekeeping genes are often used as endogenous controls. The CT-method is used when the PCR efficiency of the target and endogenous control are identical. The TaqMan probes used in paper III were designed by using Custom TaqMan® Assay Design Tool or is available directly from Life Technologies- Primer design when possible were selected to span over exon-exon boundaries to distinguish cDNA from genomics DNA and the standard curve method was used for the analysis.

Absolut Quantification by Digital droplet PCR (ddPCR)

For an absolute quantification of transcripts digital PCR can be used. In digital droplet PCR (ddPCR) the sample is portioned into 20,000 nanoliter-sized droplets enabling the measurement of thousands of independent amplification events within the single sample. In the assay some droplets will contain no templates, some will contain one and some more that one template. The number of target molecules initially presented can be determined by the number of positive and negative droplets after amplification by probability analysis.

In paper IV we used digital droplet PCR (ddPCR) (Bio-Rad). Droplets are formed by adding oil and samples (including PCR reagents, primer and probes for example TaqMan) to each cartridge, which then is placed in the QX100 Droplet Generator (Bio-Rad). The instrument generates about 20,000 droplets in a mixture of sample and oil. Then the combined one step RT and PCR is performed on a thermal cycler instrument. The PCR plate is then added to the QX100 Droplet Reader (Bio-Rad) and detection can be done by measuring fluorescence. Manual settings are used to establish a cut-off threshold for negative vs positive droplets (partitions). An absolute quantification can be performed by calculating the number of copies/µl for each sample. The analysis was performed by the TATAA center (Gothenburg).

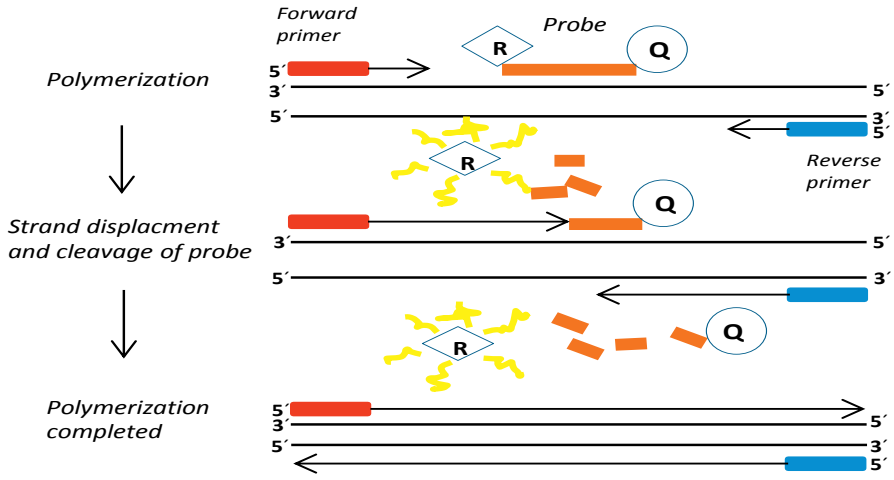

Figure 16. *Schematic principals of TaqMan analysis. Primers and probe anneals to the target DNA and under the elongation process the strand displacement and cleavage of the probe by the 5´-3´exonuclease activity is conducted. This allows fluorescence emission.*

Statistical methods

Several statistical tests have been used in this thesis. These include Student´s two-sided t-test, Fishers exact test, Analysis of variance (ANOVA) and Dunnets test.

Parametric linkage analysis

This linkage method applies to linkage analysis for disease with a known inheritance model [142]. In paper IV and V the genotype data was analyzed with the assumption of a dominant inheritance. The unit used for linkage analysis is the logarithm of odd ratio named LOD score. LOD score is the $\log_{10}$ of the likelihood of the observed genotype data when two loci are linked divided by the likelihood of the observed data when two loci are not linked. Evidence of linkage is interpreted on the total LOD score by searching for markers with a LOD score traditionally above 3. However, the LOD score that can be attained even with fully informative markers is limited by the size of the pedigree and what members are genotyped. In addition less informativity of markers gives lower LOD scores, but as long as the markers fit with the model the region is a candidate region, even for modest LOD scores. The genotype data in paper IV and V was analyzed using the software Allegro (v 2.0). A threshold of LOD 1.5 was used in selecting the regions where a disease locus might be present.

RESULTS AND DISCUSSION

Paper I

In this paper different mutation-screenings techniques including Sanger sequencing, MLPA, SSCP/HD, and PTT was used to identify mutations in the *APC* gene in patients with FAP and AFAP. Mutation screening and clinical characterization of 96 unrelated FAP patients from the Swedish Polyposis Registry revealed 61 different *APC* mutations in 81 of the 96 families and 27 of the mutations had not been reported previously. Among the characterized mutations were elusive mutations like mosaic mutations and mutations creating cryptic splice-sites. Large deletions were found in 9 % of the patients, which was higher than reported in other studies at that time. A large deletion including exon 4 of the *APC* gene could only be detected with RNA-based PTT and cDNA sequencing and not with MLPA which illustrates a limit with this otherwise high-performing technique. By using an accurate investigation of the *APC* gene with different techniques a 100% mutation detection was achieved in classical FAP patients. Real-time RT PCR revealed a lowered ASE in one FAP patient with a classical FAP phenotype, the mutation responsible for the lowered expression was identified in paper II.

Different genotype-phenotype correlations have been suggest, which are in accordance with this study, although they are not exact and differences do occur [143,144]. The most frequent found mutations were in amino acid positions 1309 and 1061 which also are in accordance with other studies. The 1309 hot-spot mutation has been found in up to 30% of population specific *APC* mutation cases [145,146]. Probands with mutations in codon 1240-1264, the main part of the MCR (mutation cluster region1284-1580) were predicted to have a severe polyposis with higher number of polyps and they were significantly younger at diagnosis compared with patients with mutations outside this region. However, the CRC risk of patients at diagnosis with mutations outside the MCR region were relatively high, explained by the high age at diagnosis for these patients.

Paper II

In paper I we reported an imbalance in allele specific expression (ASE) and a 50% reduced total gene expression of the *APC* gene in Family 1 from the Swedish Polyposis Registry. The mutation however, remained undetected and we therefore continued with the search for the disease-causing mutation. ASE of the *APC* gene as the cause of FAP has been recognized in several families and studies, but in the majority of cases no disease-causing mutation have been identified [147]. A high proportion of these patients present with classical FAP with less extra-colonic manifestations. Analysis of CNV by use of Affymetrix SNP arrays 6.0 in Family 1, revealed eventually a split deletion of around 61 kb in the upstream regulatory region of the *APC* gene, however, it was located upstream of the known major promoter 1A. Mapping of the breakpoints

was performed by using PCR, Sanger sequencing and MLPA. The mutation was found to include half of promoter 1B (320bp) and further upstream region. The deletion segregated with all the affected and none of the unaffected individuals in the family. The total expression was investigated with both TaqMan expression analysis and expression arrays, which showed a 50%-75% reduction. The ASE expression from the deleterious T allele was reduced around 90% in blood estimated from cDNA sequencing.

Since the promoter 1B was proposed to have a minor role in the regulation of the *APC* gene compared with promoter 1A, we compared the transcription from these two promoters in affected, unaffected, normal blood controls and several normal tissues. We investigated three main transcripts, two from promoter 1A and one from promoter 1B, by TaqMan analyses. The reduced expression from promoter 1B was observed in blood from mutation carriers as well as polyp tissue from the carriers. However, in blood an elevated expression from both of the transcripts from promoter 1A was found and we could not find an explanation for this. The promoter 1B was also found to be transcribed in a diversity of tissues tested and estimated in general to be higher expressed that promoter 1A.

Since our study was reported three additional studies on promoter 1B deletions have been published [148-150] (Figure 17). All of the reports present different deleted regions in one family only, except for the report by Snow et al. [149] in which a 34 kb founder mutation was detected in seven American families. The reports all present a total reduced expression from the *APC* gene at levels between 30%-70% and ASE reductions of 70%-98%, which is in concordance with our results. Even though the size of the deletions vary between 22 kb- and 132 kb all of them include the whole 1B promoter, in contrast to our, that only includes half of the promoter 1B.

In one of these studies by Pavicic et al. [148] they found the 1B promoter deletion in one family out of seven families all showing an unbalanced ASE. In the total study 51 mutation-negative FAP families were included. LOH or partial LOH was found in adenomas in the family with the 1B promoter deletion, supporting the two hit mechanism of inactivation. In their study a reduced ASE expression of only 40%-60% was observed, the remaining deleterious allelic expression seems to be optimal for tumor initiation. By exome sequencing of the same two adenomas examined by Sanger sequencing in our study (**not included in paper II**) we found a frameshift deletion in one adenoma, located between the second and third amio acid (NM_000038.5c.4192_4193del, p.Arg1399Phefs*9), in around 30% of the reads. When re-examining of the Sanger sequences, this frameshifting mutation was found to be located in the joint between two PCR fragments and therefore only visible in one direction, which could be the reason why it was not detected initially. This result, however, prove the second hit hypothesis in that sense that the wt allele with two 20-aa repeats left, in combination with a silent mutant allele (almost) will give the optimal wnt signaling. In three of these studies (including ours) absent or very few extra-colonic

manifestations has been found, which also was observed previously in families with an unbalanced ASE [101].

In the four studies the ASE expression was either measured with sequencing calculating the ratios over a SNP present in the gDNA and cDNA or by Single Nucleotide Primer Extension (SNuPE). The differences in methods used to calculate the ASE can highly influence the estimated expression levels and it is very difficult to draw any conclusion about the ASE expression in relation to different sizes of the deletions. All three studies except our show a total deletion of the *APC* promoter 1B, however, no correlation with the expression can be drawn.

By examine three transcripts from promoter 1A and promoter 1B, in different tissues we concluded that the variability of the expression from promoter 1A was higher, but the expression from this promoter seemed lower than the expression from promoter 1B in general. We therefore suggested that promoter 1B is more important than previously known. Deletion involving only promoter 1A are not common, these deletions typically affect the *APC* coding region as well. The only study with a possible 1A promoter specific deletion, was described in 2008 by Charames et al. [151], and was associated with complete silencing of the allele containing the deletion, but since the breakpoints were not mapped and they did not test for the extension of the deletion upstream of promoter 1A, it remains unclear whether or not the deletion also includes promoter1B. It is there for unclear if there exists any 1A promoter-specific deletion in FAP patients. This could be another proof of the significance of promoter 1B, as in these fours studies promoter-specific 1B promoter deletions as the cause of FAP is clearly shown proven.

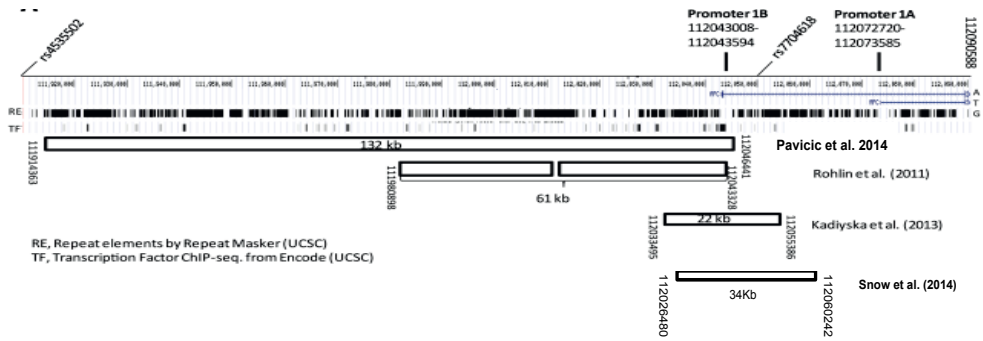

Figure 17. *Schematic overview of the four deletion found in the APC 1B promoter region ((Reprinted and modified by permission from Wiley Periodicals, Inc.: [Genes Chromosomes and Cancer] [148], © (2014).*

In mutation carriers with lowered level of expression from promoter 1B, we did found elevated levels of expression from promoter 1A, this was only observed in blood, and not in normal colon mucosa or adenoma from mutations carriers. The reason for this could be tissue specific compensation due to damaging promoter 1B. None of the other three studies have examined transcripts from the two promoters. In the last study by Snow et al [149], they found gastric and duodenal polys, speculating that the loss of expression from *APC* promoter 1B in combination with low levels of promoter 1A expression might modify the presence of gastric and duodenal polyps. However the combinatory regulation of the *APC* gene by promoter 1A and promoter 1B requires further investigations.

Paper III

In inherited diseases with *de novo* mutations, mosaic mutations can be present. Mosaic mutations can have important consequences for the patients and their families. Identification of mosaic mutations is technically challenging but crucial. As the first platform for MPS (454 from Roche) was introduced, we evaluated this new technique together with the main mutation-detection techniques used at that time for their ability to detect mosaic mutations. Two known mosaic mutations without confirmed mosaic frequency were used together with dilution to known concentrations of 7 other known mutations to access the sensitivity of the methods. The known mosaic mutations were, *APC* (c.2700_2701 delTC) detected only by SSCP in paper I and the second one was the *NF2* mutation (c.1026_1027delGA) detected by Sanger sequencing and not previously reported. .

In this study we calculated the mosaic-allele frequency based on the mutant allele as had been done in previous similar studies at that time [20,21]. Dilutions were made from a known "full" heterozygote mutation and the undiluted sample was set to be 100% mutant allele. However, in recent studies using MPS, the frequency of mosaic mutations are generally based on the total read count. To compare our data with these more recent studies we need to divide our frequency results by two.

Regarding conventional methods we found SSCP/HD and DHPLC to be the most sensitive methods, where DHPLC were slightly more sensitive regarding substitutions and SSCP more sensitive regarding insertion and deletions. PTT was also found to be sensitive, but only three of the samples were included in the analysis and conclusions from this are difficult to make. However, even recent studies with PTT analysis suggest it to be a very sensitive method as it has been shown to be able to detect frequencies down to 3% of mutant alleles. insertion /deletions) [152]. Sanger sequencing, which has long been used as the golden standard in mutation detection of sequence variants, could not detect mosaic mutations below 15 % (the lowest detection level in regards to mutant allele only in this study). It is difficult or almost impossible to detect mosaic mutations below 10-15% with this method especially insertion/duplication showed a low sensitivity with this method.

Technically-based challenges in the analyses were variation in concentration depending on pipetting errors and difficulties quantifying the amplicons accurately. The use of a Taq polymerase without proof-reading activity certainly also affected the results as incorrect bases can be inserted in the PCR amplification. This was also reflected in the mis-call frequency (which reflects the most abundant false-positive variant found in each position) which was 0.05%-5%, where also the coverage was variable.

This study was done in the early MPS days and also data analyses were a great challenge. The results however, showed that MPS can detect mutations at low frequencies regardless of type of mosaic mutation like the *APC* c. 2700_2701delTC which was present in 4.5% in DNA from blood leucocytes and which could not be detected by Sanger sequencing (Figure 18).

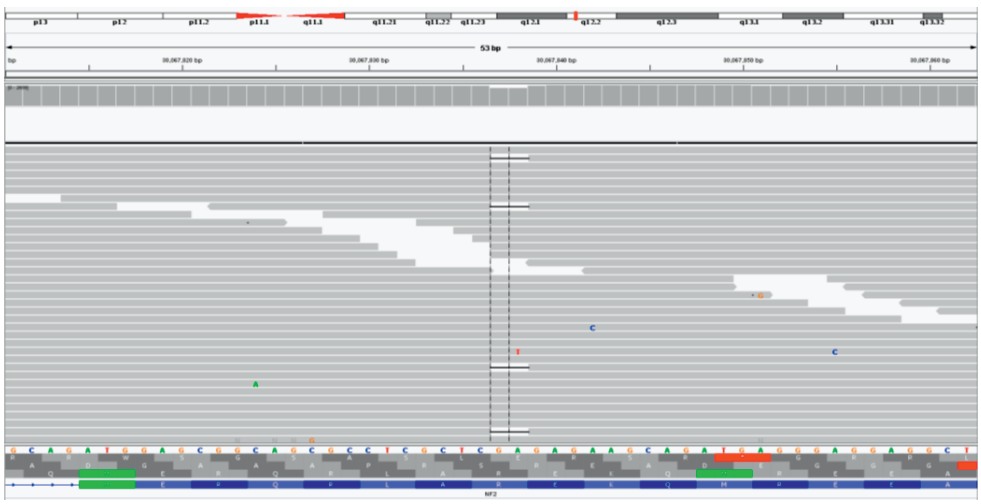

Figure 18. *Overview of the NF2 gene region were the mosaic mutation c.1026_1027delGA is present in 14% of the reads when analyzing DNA from blood leucocytes. Reads containing the deletion is illustrated with a black line.*

Using ultra-deep sequencing (>1000x coverage [133]) mosaic mutations have recently been identified in the *PTEN, TP53* and *VHL* genes [23,25,153,154]. Somatic mosaic mutations in *de novo* cases as well as gonodal mosaicism in parents of disease affected children was presented, emphasing the importance of this method in clinical genetic analyses. The depth of sequencing has varied between different studies. In our study a cut off was set to detect 50 variant alleles to be able to get a high confidence. To detect a variation at 1% a sequencing depth of 5,000 where then needed. The statistical confidence for 1% mosaic mutation detection with the 454 sequencing system was in a study by Izawa et al. set to 700 x coverage [155]. The 454 sequencing system was also used to detect mosaic mutations in *VHL* down to 1.7 % (68 mutant read/4,059 total reads). Studies using Illumina sequencing system with different targeted approaches

detects low frequency mutation down to around 1% as well [154]. In our study a detection frequency down to 1% was reached.

Paper IV and V

In paper IV and V exome sequencing has been used to search for new possible disease-causing variants. This might be difficult when dealing with a large number of variants that are called in each exome. The value of being able to compare data from affected with unaffected individuals from the same family, controls from e.g. in-house datasets and large common databases is crucial. "Private" variants common in the family in both affected and unaffected as well as highly polymorphic variants and platform specific artifact can be filtered, leaving fewer variants for further evaluation.

In addition to segregation analyses and in-house controls we used linkage analysis to narrow down the search region in Family C (paper IV). This however left us with the opposite problem when only one variant remained after the filtration This was a variant in the *TRPM1* gene, which was considered to be an unlikely candidate because it had a MAF of 0.007, was present in our in-house controls and had not been associated with CRC before. No other very likely candidates were called in the non-coding region in the sequencing either.

By additional manual reviewing of variants in the area of interest, a variant that had been missed by the caller the *GREM1* c.-76 C>G variant was found. It was a CpG-site located in the promoter CpG island of *GREM1* and was found to segregate with affected but not unaffected individuals in the family and the variant was unknown in the general population and not present in our in-house data set. Bisulfite sequencing of the whole CpG island region (including c.-76 C>G), revealed no difference in methylation pattern between affected, unaffected and controls, in fact the c.-76 position was not significantly methylated in any of the samples. No disease-causing epigentic mechanism could be proposed.

A large duplication in the upstream region of the *GREM1* had recently been published as the disease-causing mutation in HMPS [94]. We analyzed the region in our family for SVs by using MLPA. With this method we identified a duplication of a region upstream of the *GREM1* gene in Family C. This region included an enhancer element that had previously been found to cause an up regulated and ectopic expression of the *GREM1* gene [94]. Our mutation included around half of the duplicated region found in the reported family and the enhancer element was remaining. We performed an expression analysis of normal colon epithelium available from two patients and seven control sample with digital droplet PCR and found a significantly higher expression of the *GREM1* gene also in Family C. The phenotype of the family was initially considered to be AFAP, however, as juvenile-like polyps were found in some affected family members, a mixed polyposis phenotype was proposed. This duplication of around 16Kb is considered to be the disease-causing mutation in the Family C.

GREM1 and proposed disease mechanism

The colonic surface epithelium is composed of columnar cells in as single layer and goblet cells. It also has crypts mainly composed of goblet cells, but with a few undifferentiated progenitor or stem cells at the base. The stem cells undergo mitotic division and migrate towards the top of the crypts [156]. Human colonic epithelial-cell renewal, proliferation and differentiation are stringently controlled by numerous regulatory pathways, including the Bone morphogenetic proteins (BMPs). Intestinal epithelial stem cells are also supported by underlying intestinal subepithelial myofibroblasts (ISEMFs), which are in close connection with smooth muscle cells. These cells are at the base of the intestinal crypts and may contribute to the stem cell niche acting as regulators of intestinal stem cell self-renewal and differentiation. There is a difference in genes expressed in the colon crypts and tops. BMP antagonist *GREM1* and *GREM2* are expressed in ISEMFs and smooth muscle cells (SMC) at the colon crypts. BMP signaling pathway counteracts Wnt signaling toward the top of the crypts allowing for differentiation. GREM1 further activates Wnt signaling in epithelial cells, a model for the expression of Wnt signaling and BMP signaling in colon crypt base and top [157-159].is illustrated in Figure 19.

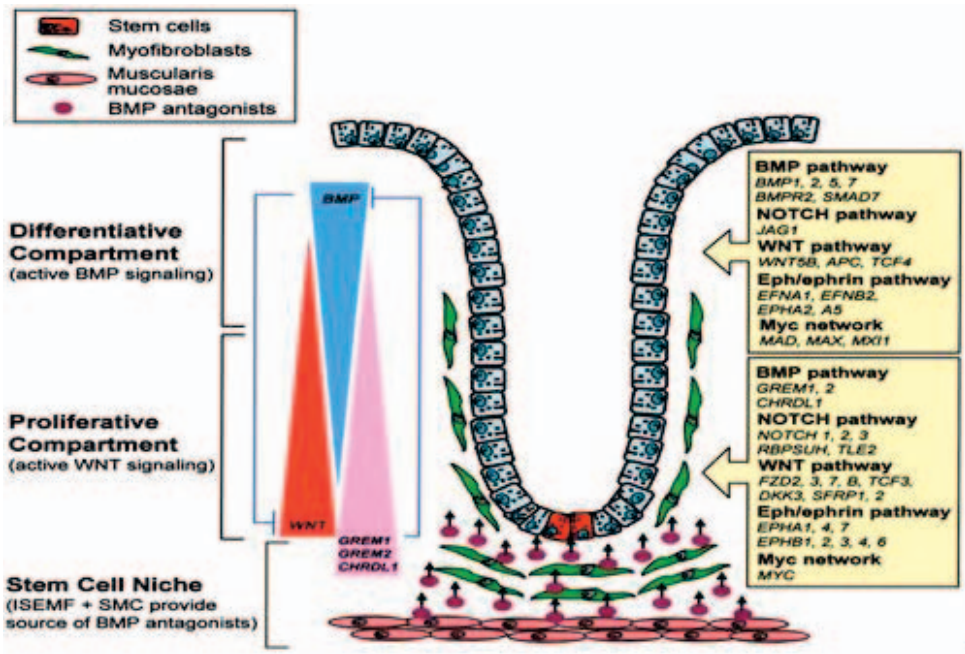

Figure 19. *A schematic view of the pathway expressed in the colon crypt base and top. A counteracting relation between the BMP signaling and Wnt-signaling and the correlation with GREM1 and GREM2 expression is illustrated (Reprinted with permission from National Academy of Sciences, U.S.A © 2007, from [157]).*

Jaeger et al [94] examined the localization of the *GREM1* expression in normal colorectal crypts of individuals in their family with the duplication and those without and in controls. In controls and family members without the duplication the expression was restricted to the intestinal subepithelial myofibroblasts (ISEMFs) at the crypts base. However, in individuals with the duplication the expression of *GREM1* was not only expressed at the basal ISEMFs, but also at very high levels in the epithelial cells extending up the sides of the crypt. This strong overexpression in the intestinal epithelium was defined as the disease mechanism. As our family also presented an up-regulated expression of the *GREM1* gene in the colon epithelium, we suggest that this is the disease mechanism in Family C as well.

In paper IV two additional families, Family A and B, were analyzed by exome sequencing, however, as only two affected individuals from each family were available and no unaffected it was more difficult to suggest possible disease causing candidate variants in the analyses. We therefore only considered genes with truncating variants and genes harboring none-synonomous variants that had previously been found having known relationships with pathways, and biological processes, in colorectal cancer. When considering truncating variants, LoFs can be present in healthy individuals as reported by McArthur et al. [3,29]. None of the truncating variants that were found in this study was found to be a LoF in healthy individuals i.e. a polymorphism. Genes of interest were sequenced in 107 additional patients (without previously detected disease-causing mutations in CRC predisposing genes).

From these analyses rare truncating variants were detected in *LMO7*, *GPLD1*, *EXT2*, *FBXL13*, *CLCA4*, *ECT2L*, *MMP8* and *LIG4* in Families A and B. In addition three missense variants were identified by the filter criteria used; *TCF7*, *RET* and *BUB1B*. No truncating variants, but possible deleterious amino-acid substitution variants were found in *BUB1B*, *LMO7* and *MMP8* in the additional 107 patients sequenced. The last three genes harboring both truncating variants and possible deleterious amino acids and might therefor have a broader interest in CRC predisposition. The genes have been considered to be tumor suppressor genes and variants/mutations in these genes have previously also been associated with CRC. Germline mutations in *BUB1B* have recently been found leading to increased risk for gastrointestinal tumors and predispose to increased risk of colorectal cancer at a young age. Linkage to the *LMO7* gene locus region was recently identified in a large CRC family from Utah presenting a few adenomas. These genes might therefore present interesting candidates in hereditary CRC. When we performed the exome sequencing we limited the analyses to examine truncating variants in Families A and B, today this would not have been done in this way, today all variants in the whole exome analysis had to be considered.

In paper V we identified a disease-causing mutation in the *POLE* gene c.1089C>A, p.Asn363Lys, by using the same approach as in paper IV, but without the linkage analysis included from the start. The *POLE* variant was selected as a possible candidate, even though it was an amino-acid substitution, based on the involvement in the DNA repair pathway. The mutation was confirmed based on the segregation analysis in this

large family. Evaluation by local normal controls and common databases was done as well. In silico functional amino-acid prediction suggest a deleterious effect of the substitution and the amino-acid substitution was also suggested to have a profound effect on the substrate binding at the active site of the proofreading exonuclease domain based on a structure of a yeast POLE protein. The amino acid was also highly conserved between species. High penetrant mutations in *POLE* and *POLD*1 were recently published by Palles et al [97] confirming these genes to be mutated in CRC syndromes.

Palles et al [97] screened 3805 CRC patients and found the p.Leu424Val variant in 12 patients and the *POLD1* p.Ser478Asn in three patients with multiple or very large adenomas or multiple colorectal carcinomas or early onset CRC [97]. In addition they found another *POLD*1 variant c.981C>G, (Pro327Leu), which were considered pathogenic. All *POLD1* mutation carriers also displayed endometrial cancer whereas the *POLE* carries only had colorectal cancer. However our family presented a wide tumor spectra including ovarian, pancreatic and brain tumors. Additional studies by Valle et al. [160] detected a single de novo *POLE* p.Leu424Val mutation in a polyposis patient with 35 adenomas, hyperplastic and mixed polyps. They also found a new *POLD*1 potential mutation (Leu474Pro) in a non-polyposis patient with MSS and with endometrial cancer by screening of 858 families with early onset CRC and polyposis. To complicate the picture further, germline *POLE* p.Leu424Val was recently also found in two families with MSI tumors. The immumohistochemistry analyses showed deficiency of MSH6/MSH2, but no germ-line mutation in the MMR genes was identified. By sequencing of tumor DNA from one patients from each family two somatic mutations in *MSH6* and two somatic mutations in *MSH2* were identified in each tumor [161]. All these studies suggest that disease-casuing mutations in these genes are rare and that they are located in the exonucleas domain. Seemingly they are present in both polyposis and non-polyposis patients and no consisted genotype-phenotype correlations can be concluded.

In paper IV we did also find a missense mutation in the *POLE* gene c.1274A>G, p.Lys425Arg in an early onset CRC non-polyposis patient without any family history. New proposed disease-causing missense mutations in the exonuclease domains of *POLE* and *POLD1* are now being published, mainly in single families [160]. These genes are now considered being included in routine diagnostic panels, however, careful consideration has to be taken when classifying these variants as pathogenic and functional analysis, as well as familial segregation, is desirable.

Exome sequencing in familial early onset CRC using a large number of independent cases, aimed at identifying rare coding variants of varying penetrance, have been conducted in several studies. This approach has however been unsuccessful for identification of new common predisposing genes shared among many families, in this cases single family analyses have been shown to be a better approach. When Palles et al [97] employed whole genome sequencing in several single cases of AFAP families without identified mutation this approach failed to identify candidate genes shared

among 4≥ individuals. When they instead analyzed several individuals from the same family in a single family approach, they identified the *POLE* and *POLD1* mutations [97]. In FCCX a mutation in RPS20 and a mutation in *SEMA4A* were also identified with a similar single family approach [120,121].

Paper VI

In this paper we have analyzed 67 anonymized CRC patients without detected disease-causing mutations in a panel consisting of 19 CRC susceptibility genes, including well-defined syndrome-specific genes. The genes included were *APC, MUTYH, BMPR1A, SMAD4, STK11, PTEN, AXIN2, CTNNB1, MLH1, MSH2, EPCAM, MSH6, PMS2, MLH3, MSH3, PMS1, CDH1, MET, CHEK2*. Regions 50 kb upstream and downstream of all genes and all intronic regions were included in the target regions except for the *APC* gene for which 100 kb upstream was included and for the MET gene only coding exons were targeted. The patients were divided into clinical subgroups and the panel and pipeline were validated by positive controls. The purpose of the study was to identify causative mutations in polyposis and non-polyposis patients without previously identified mutations by syndrome-specific screenings. The purpose was also to classify detected variants by available *in silico* prediction tools, databases and in-house information of local polymorphisms.

Seven truncating and splice site mutations (nonsense or frame shift) were found and three class 5 mutations. This gives a mutation detection frequency of 15% (10/67) in the 67 index patients analyzed. 12% (8/67) were found in clinical actionable genes, including; *APC, MUTYH, BMPR1A, SMAD4, STK11, PTEN MLH1, MSH2, MSH6, EPCAM* and *CDH1*. In addition, two structural variants were found, one deletion (hg19 chr18:48537165_48539080del) around 1.9 kb located 2 kb upstream of the *SMAD4* gene and a *CDH1* duplication in intron 1 (hg19/ Chr16:68802080_68826280dup) which was around 24.2 kb in size.

Traditionally mutations in *BMPR1A* and *SMAD4* are associated with the juvenile polyposis syndrome. In this study three *BMPR1A* and one *SMAD4* mutation were found in patients who presented with FAP, mixed polyposis or early onset CRC without polyps. This highlights the complexity of the genotype to phenotype correlations in these genes. This can also be related to the overlap in polyp histotypes as illustrated in Figure 20. Mutations in *BMPR1A* and *SMAD4* have recently been found in an extended phenotypic spectrum beyond juvenile polyposis, HMPS, FAP, AFAP, FCCX and early onset CRC without familial history with MSS tumors [162]. There is a complexity of possible ligand receptor and downstream effector combinations in the BMP/TGFR-β signaling pathways and this might explain part of the genotype-phenotype relationship. There might also be a genotype-phenotype correlation depending on where in the gene the mutation is located.

In the *SMAD4* deleted region upstream of the gene an insulator element was found. An insulator element can act as a barrier to enhancer action and transcription of genes

beyond the insulator and is not stimulated by the enhancer when the insulator is active. This deletion might therefore have an effect on the expression of the gene. The deletion might lead to an up regulated expression of the *SMAD4* gene, which then presents a gain of function effect, however, further expression studies are needed in order to draw any conclusions.

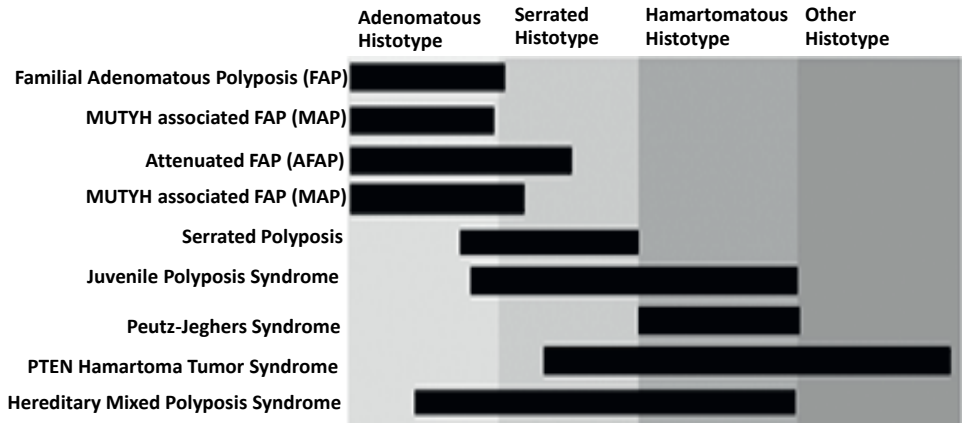

Figure 20. *Relationships and overlaps of the different polyp histotypes with polyposis phenotypes (Reprinted with permission from Wiley Periodicals, Inc , © 2013)[92].*

In this study we have also manually tried to classify missense variants in cases without an InSiGHT classification. By using publications, locus-specific databases and HGMD as well as in-silico predication tools we have tried to specify as similar scale as InSiGHT for classification. In Table 1 we have specified some of the criteria used. We have also considered in-house information to be able to find local polymorphisms or risk alleles and effects on splicing for all variants as well.

By classification of variants by consortiums like InSiGHT, the use of national and international control databases as well as more functional analysis available on allelic variants will make classification easier. Variable penetrance and modifier gene variants will however still be a challenge.

Table 1

Missense classification criteria	
Class 5- Pathogenic	Classified as pathogenic according to publications, LOVD and HGMD, predicted damaging according to in silico protein prediction tools, evolutionarly conservarion of nucleotides and seggregation analysis in family.
Class 4 -Likely pathogenic	Classified as pathogenic or probably pathogenic according to publication, LOVD and HGMD, predicted damaging or probabaly dagmaging according to in silico protein prediction tools and evolutionarly conservation of nucleotides.
Class 3 -Uncertain	Insufficient evidence to classify as class 1, 2, 4, or 5
Class 2-Likely not pathogenic	Variants reported to occur in the general population at a frequency < 1%, and has not been reported as pathogenic according to publications, LOVD and HGMD and are predicted benign or probaly benign accoring too in silico protein prediction tools.
Class 1-Not pathogenic	Variants reported to occur in the general population at frequency ≥1%.

In 7.4% of patients a VOUS was found and the majorities were located in Lynch-associated genes in concordance also with other reports. We did also find a possible risk allele in the *CHEK2* gene. *CHEK2* is a protein kinase activated in response to DNA damage and involved in cell cycle arrest. It is a tumor suppressor gene and mutations result in decreased DNA repair activity or an inability of the cell to undergo apoptosis. A known risk allele c.470T>C, p.Ille157Thr has been found in CRC and we found this variant in a patients with AFAP. In addition we did found a splice variant, c.319+2T>A, in a patient with familial CRC, polyps and a MSS tumor, this might also possible present a risk allele. *CHEK2* risk alleles have also been found in several other panel studies. No diagnostic criteria or guidelines are yet available for mutations in *CHEK2* in CRC.

Our results are in concordance with the results obtained in other studies of similar gene panels [162,163]. In a recent panel study of 557 colon patients 30% of the probands with mutations in well-defined genes did not meet the clinical criteria for the associated syndrome [164], in our study this was at least 37% (3/8). By use of a panel of known CRC susceptibility genes for a broader screening of hereditary CRC syndromes the disease-causing mutation can be identified to a higher extent. This will lead to improved mutation detection analysis for diagnostic and carrier testing.

CONCLUSIONS AND FUTURE PERSPECTIVE

The main purpose of our studies has been to be able to develop and provide high quality mutation detection analyses by which we can identify causative mutations in patients with CRC syndromes. The future aim is also that the mutations can be used for personalized surveillance and treatment for the patient.

In paper I we achieved a 100% detection frequency in classical FAP by using a variety of mutation detection techniques. In paper II we introduced the SNP array method, which also made it possible to identify and characterize the mutation in the promoter regions of *APC* previously not covered by the methods used. Around 20% of FAP patients are thought to harbor mosaic mutations, which can also be true for other familial cancer syndromes. To be able to evaluate the possibilities to detect these mutations in a diagnostic genetic setting we conducted a sensitivity study (paper III) in which we included the MPS method.

By using whole exome sequencing (WES) and targeted gene sequencing we have then demonstrated the possibilities in identifying new disease-causing genes in familial colorectal cancer syndromes as well as identifying mutations in known familial colorectal cancer genes not correlated with the considered phenotypic presentation (paper IV-VI). This will improve diagnostics and carrier testing for families with inherited colorectal cancer syndromes.

The MPS method has made it possible to identify new disease-causing genes in hereditary syndromes mainly by using WES. With WES mutations outside exons cannot be identified, which was experienced and shown in paper IV. With reducing costs of WGS extended possibilities of mutation detection will be available and by using the PCR free methods artifacts introduced in the PCR step will be abolished. Several other recent updates have emerged as well, using instruments like the PACBIO RSII (Pacific Biosciences), sequencing of long reads (>2kb) is now possible. This sequencing technique allows for e.g. phasing of variants to determine the phase (meaning analyzing if variants are on the same or different alleles), resolve structural variations with better confidence, isoform detection from RNA sequencing and many more.

Finding new cancer predisposition genes will be of importance, not only for the families, but also for improved understanding of the genetic factors underlying tumor development in general. By investigating the effect each mutation have on the genes, the protein product and the function of the cell this will hopefully contribute to the major sequencing efforts being conducted to improve the diagnostics and treatment of patients.

POPULÄRVETENSKAPLIG SAMMANFATTNING

Kolorektal cancer (tjock- och ändtarmscancer) är en av de vanligaste cancerformerna, med en miljon nya fall i världen och cirka 4 000 nya fall i Sverige varje år. Det är också en cancerform som till stor del beror på ärftliga faktorer, uppskattningsvis till 20-30 %. Personer med ärftlig cancer får den i regel tidigare i livet och löper också större risk att drabbas av fler än en tumör, än personer med sporadisk (icke ärftlig) cancer.

5-6 % av de ärftliga cancerformerna kan delas in i syndrom med mutationer i kända gener. Den vanligaste typen av ärftlig kolorektal cancer är Lynch Syndrom, som står för upp till 3 % av all kolorektal cancer. Den näst vanligaste är Familjär Adenomatös Polypos (FAP), som står för cirka 1 %. FAP orsakas främst av mutationer i APC-genen och ärvs dominant.

Normalt finns alla gener i två kopior. Vid dominant nedärvning räcker det att ärva en gen med mutation från den ena av sina föräldrar, för att löpa risk att få sjukdomen. I cellerna är det däremot en recessiv sjukdom, vilket innebär att den normala genkopian, som ärvts av den andra föräldern, måste slås ut i en cell innan sjukdomen kan uppstå. De flesta av dessa syndrom har en dominant nedärvning.

I den klassiska formen av FAP med hundra till tusentals polyper i tjock- och ändtarm, hittas nästan alla mutationer i APC-genen. Men i den mildare formen av FAP, så kallad Attenuerad FAP (AFAP) med färre än 100 polyper, hittas endast cirka 30 % av mutationerna i APC-genen. Det finns alltså en stor andel av ärftlig cancer där den sjukdomsorsakande mutationen ännu inte har kunnat identifieras.

Avhandlingens mål

Målet var dels att kartlägga mutationerna i APC-genen hos patienter med FAP, dels att med nya tekniker leta efter ovanliga mutationer i APC-genen. Målet var också att i familjer med en atypisk sjukdomsbild och i familjer med en mildare FAP, leta efter nya gener med mutation.

Arbetsmetod och resultat

Den **första studien** gjordes på polypos-patienter från det svenska polyposregistret. Genom att använda ett flertal kompletterande tekniker, kunde vi detektera mutationer i APC-genen i hela 100 % av familjer med klassisk FAP.

I en av familjerna – den största FAP-familjen i Sverige (cirka 150 personer, varav 59 med verifierad FAP) – hittades ett lågt uttryck av genen, men den faktiska mutationen bakom detta kunde inte identifieras.

I den **andra studien** letade vi efter mutationen som orsakade det låga uttrycket. Här använde vi oss av SNP Array-tekniken, en teknik som ger möjlighet att hitta stora förändringar av regioner i hela genomet. Vi kunde då konstatera ett bortfall av en del av promotorregionen i APC-genen. Bortfallet är den sjukdomsorsakande mutationen. I många familjer har man tidigare sett det lägre uttrycket av APC-genen, men inte kunnat hitta den orsakande mutationen till detta. Vår studie var den första där man hittat den orsakande mutationen.

I den första studien hade vi också identifierat en ovanlig mutation; en mutation som inte fanns i alla cellerna i blodet, en så kallad mosaikmutation. Dessa mutationer är svåra att hitta och i **Studie III** gjorde vi därför en utvärdering med hjälp av fyra tekniker, inklusive Massiv Parallell Sekvensering. Detta gjorde det möjligt att på ett helt nytt sätt kunna identifiera mosaikmutationerna.

Ny revolutionerande metod

Kartläggningen av det mänskliga genomet och introduktionen av Massiv Parallell Sekvensering, även kallad Next Generation Sequencing, har revolutionerat utvecklingen under de senaste åren. Denna teknik gör det möjligt att idag sekvensera alla kodande genregioner, till och med hela genomet, snabbt och till en låg kostnad. Metoden ger oss betydligt större möjligheter än tidigare att identifiera gener med sjukdomsorsakande mutationer.

I **Studie IV och V** har vi använt hos av sekvensering av alla kodande genregioner för att leta efter gener, som kan innehålla den sjukdomsorsakande mutationen.

- I Studie IV kunde vi identifiera den sjukdomsorsakande mutationen i genen GREM1 i en familj. Mutationen kunde identifieras i en familj med AFAP, men som även hade lite speciella typer av polyper.
- I studie V identifierade vi genen POLE. Mutationen hittades i en stor familj med ett brett tumörspektrum, inkluderande tumörer i äggstockar, livmoder och hjärna. Familjen har ingått i forskningsstudier under 30 år.

Massiv Parallell Sekvensering gör det möjligt att sekvensera alla våra cirka 23 000 gener samtidigt, istället för endast en eller ett fåtal samtidigt. Metoden ger oss även en stor mängd data, som man kan studera och sålla bland. Genom att exempelvis studera om variationerna är vanliga, så kallade polymorfier som finns hos de flesta individer, kan man helt avskriva en del. Därefter kan man göra ytterligare jämförelser i olika databaser och även analysera friska och sjuka familjemedlemmars DNA för att hitta de mutationer som kan orsaka sjukdom.

I **Studie VI** begränsade vi sökningen till en panel av gener som tidigare är kända för att orsaka kolorektal cancer. I denna studie kunde vi då se att presentationen av sjukdomen inte alltid stämmer överens med de gener där man traditionellt tror att mutationen

borde finnas. Slutsatsen är att överlappningen och komplexiteten mellan de olika syndromen alltså är större än vad man tidigare trott.

Resultatet av studierna

- I våra studier har vi framför allt identifierat nya gener med mutation med hjälp av den nya massiva parallellsekvenseringstekniken.
- Vi har även kunnat kartlägga komplexa svårdetekterade mutationer i kända gener förknippade med ärftlig kolorektal cancer.
- Vi kan fastslå att en bredare testning behövs av kända gener för att identifiera den sjukdomsorsakande mutationen.

Dessa resultat leder till en ökad diagnostik av familjer och individer med ärftlig kolorectal cancer.

Dessa kunskaper kan också leda till en ökad förståelse för:

- Vilka individer som har ökad risk att utveckla kolorektal cancer.
- Vilka genetiska mekanismer som orsakar att kolontumörer uppstår och hur de progredierar.
- Hur dessa genetiska mekanismer kan användas som prognostiska markörer för att förbättra den personbaserade behandlingen vid såväl ärftlig som sporadisk kolorectal cancer.

ACKNOWLEDGEMENTS

Stort tack till!

Alla patienter och familjer som ställt upp på att delta i forskningsstudierna. Jag hoppas att de kunskaper vi därigenom vunnit kommer er till godo.

Alla tidigare och **nuvarande** kollegor på Klinisk Genetik som genom er kompetens, ert engagemang och glädje bidragit till en utvecklande, varm och familjär och arbetsplats.

Min huvudhandledare **Margareta Nordling** för ditt genuina forsknings engagemang och din stora kompetens inom familjär cancer, du vill alltid göra det som är bäst för patienterna. Vi delar ett stort intresse för nya tekniker och jag är oerhört glad att jag har fått möjlighet att jobba med detta och att du inte (tror jag) har sagt nej till någon av mina idéer som jag har velat genomföra. Vi har också kompletterat varandra bra och därför kunnat genomföra ganska mycket trots att vi är en liten forskningsgrupp!

Min bihandledare **Jan Björk** för att du delat med dig av dina kliniska kunskaper och att jag fick göra studiebesök på polypsregister

Min bihandledare **Staffan Nilsson** för din oumbärliga statistik hjälp

Till **alla mina labvänner på plan 4** för att ni har stöttat mig i alla år på privat och i jobbet. Utan er hade jag inte stått här idag **Maria Badenfors**, **Eva Lotta Kärrstedt**, **Maria Yhr**, **Eva Portinsson**, **Julia Rundberg**, **Mirja Marcher**, **Josephine Wernersson också** för all hjälp med forskningen**, Yvonne Engvall** för att ha introducerat mig i alla molekylärbiologiska metoder. Din fantastiska tekniska kompetens gör att du har en förmåga att få de mest omöjliga experiment att fungera. **Frida Eiengård** som har varit en klippa inom projektet de senaste åren och bidragit fantastiskt mycket till att denna forskning har kunna genomföras.

Birgitta Hallberg för ditt stora forskningsengagemang du har lagt ner, speciellt för familjen M. Din insats har varit oumbärlig.

Torbjörn Olausson som stått ut med mig som rumskompis och som alltid är hjälpsam!

Anders Kvist för att du svarat på alla mina möjliga och omöjliga bioinformatiska frågor. Tack även till **Eva Rambech**, **Therese Törngren** och **Åke Borg,** det har varit roligt att sammarbeta med er.

All personal på Genomics och Bioinformatics, speciellt tack till **Marcela Davila** och **Lisa Olsson** för att ni alltid är hjälpsamma och engagerade.

Ulf Lundstam och **Theofanis Zagoras** för all sammaställning av klinisk data

Alla på plan 3 där jag har suttit det senaste året, **Rose-Marie Sjöberg** som man tillsammans med alltid råkar ut för nya oanade äventyr, t ex ”stenresan” i Utah, **Ingrid Eriksson** tack för trevliga morgonsamtal som har hjälpt mig när jag har skrivet denna bok.

För allt stöd från familj och vänner utanför forskningen, mina föräldrar **Lena** och **Claes-Göran**, syskon, mina svärföräldrar **Lena** och **Lennart** och alla andra familjemedlemmar. **Anna Lundgren** för att du är ”inom samma bransch” och den enda som förstår mig! Alla morgonpromenader i Espevik, en ventil då vi kan diskutera allt hög och lågt.

Rigmor och **Lasse** för att ni alltid stöttar mig

Mormor **Inga**, min mentor som jag önskar hade varit med mig idag.

Slutligen min familj, mina barn **Elsa** och **Nils** tack för att ni finns det är fantastiskt att se världen genom era barnögon, **Rickard** tack för ditt stöd, speciellt din insats under denna period. Utan dig hade denna avhandling inte blivit skriven. Jag älskar er!

REFERENCES

1. Watson JD, Crick FH (1953) The structure of DNA. Cold Spring Harb Symp Quant Biol 18: 123-131.
2. Hastings ML, Krainer AR (2001) Pre-mRNA splicing in the new millennium. Curr Opin Cell Biol 13: 302-309.
3. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823-828.
4. Thusberg J, Olatubosun A, Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32: 358-368.
5. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, et al. (2010) A method and server for predicting damaging missense mutations. Nat Methods 7: 248-249.
6. Gonzalez-Perez A, Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88: 440-449.
7. Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M (2012) PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33: 1166-1174.
8. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.
9. Cooper GM, Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat Rev Genet 12: 628-640.
10. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM (2014) A general framework for estimating the relative pathogenicity of human genetic variants. 46: 310-315.
11. Thompson BA, Greenblatt MS, Vallee MP, Herkert JC, Tessereau C, et al. (2013) Calibration of multiple in silico tools for predicting pathogenicity of mismatch repair gene missense substitutions. Hum Mutat 34: 255-265.
12. Thompson BA, Spurdle AB, Plazzer JP, Greenblatt MS, Akagi K, et al. (2014) Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet 46: 107-115.
13. Corvelo A, Hallegger M, Smith CW, Eyras E (2010) Genome-wide association between branch point properties and alternative splicing. PLoS Comput Biol 6: e1001016.
14. Taggart AJ, DeSimone AM, Shih JS, Filloux ME, Fairbrother WG (2012) Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. Nat Struct Mol Biol 19: 719-721.
15. Kanter-Smoler G, Fritzell K, Rohlin A, Engwall Y, Hallberg B, et al. (2008) Clinical characterization and the mutation spectrum in Swedish adenomatous polyposis families. BMC Med 6: 10.
16. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12: 363-376.
17. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, et al. (2010) Diversity of human copy number variation and multicopy genes. Science 330: 641-646.

18. Biesecker LG, Spinner NB (2013) A genomic view of mosaicism and human disease. Nat Rev Genet 14: 307-320.
19. Evans DG, Bowers N, Huson SM, Wallace A (2013) Mutation type and position varies between mosaic and inherited NF2 and correlates with disease severity. Clin Genet 83: 594-595.
20. Aretz S, Stienen D, Friedrichs N, Stemmler S, Uhlhaas S, et al. (2007) Somatic APC mosaicism: a frequent cause of familial adenomatous polyposis (FAP). Hum Mutat 28: 985-992.
21. Hes FJ, Nielsen M, Bik EC, Konvalinka D, Wijnen JT, et al. (2008) Somatic APC mosaicism: an underestimated cause of polyposis coli. Gut 57: 71-76.
22. Schwab AL, Tuohy TM, Condie M, Neklason DW, Burt RW (2008) Gonadal mosaicism and familial adenomatous polyposis. Fam Cancer 7: 173-177.
23. Gammon A, Jasperson K, Pilarski R, Prior T, Kuwada S (2013) PTEN mosaicism with features of Cowden syndrome. Clin Genet 84: 593-595.
24. Salo-Mullen EE, Shia J, Brownell I, Allen P, Girotra M, et al. (2014) Mosaic partial deletion of the PTEN gene in a patient with Cowden syndrome. Fam Cancer.
25. Behjati S, Maschietto M, Williams RD, Side L, Hubank M, et al. (2014) A pathogenic mosaic TP53 mutation in two germ layers detected by next generation sequencing. PLoS One 9: e96531.
26. Ruark E, Snape K, Humburg P, Loveday C, Bajrami I, et al. (2013) Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. Nature 493: 406-410.
27. van Arensbergen J, van Steensel B, Bussemaker HJ (2014) In search of the determinants of enhancer-promoter interaction specificity. Trends Cell Biol.
28. Chen J, Weiss WA (2014) When deletions gain functions: commandeering epigenetic mechanisms. Cancer Cell 26: 160-161.
29. MacArthur DG, Manolio TA, Dimmock DP, Rehm HL, Shendure J, et al. (2014) Guidelines for investigating causality of sequence variants in human disease. Nature 508: 469-476.
30. Condit CM, Achter PJ, Lauer I, Sefcovic E (2002) The changing meanings of "mutation:" A contextualized study of public discourse. Hum Mutat 19: 69-75.
31. Howard HJ, Horaitis O, Cotton RG, Vihinen M, Dalgleish R, et al. (2010) The Human Variome Project (HVP) 2009 Forum "Towards Establishing Standards". Hum Mutat 31: 366-367.
32. Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, et al. (2007) The genomic landscapes of human breast and colorectal cancers. Science 318: 1108-1113.
33. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., et al. (2013) Cancer genome landscapes. Science 339: 1546-1558.
34. Stratton MR, Campbell PJ, Futreal PA (2009) The cancer genome. Nature 458: 719-724.
35. Knudson AG, Jr. (1971) Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68: 820-823.
36. Kinzler KW, Vogelstein B (1997) Cancer-susceptibility genes. Gatekeepers and caretakers. Nature 386: 761, 763.
37. Santoro M, Carlomagno F (2013) Central role of RET in thyroid cancer. Cold Spring Harb Perspect Biol 5: a009233.
38. Luo Y, Tsuchiya KD, Il Park D, Fausel R, Kanngurn S, et al. (2013) RET is a potential tumor suppressor gene in colorectal cancer. Oncogene 32: 2037-2047.

39. Haramis AP, Begthel H, van den Born M, van Es J, Jonkheer S, et al. (2004) De novo crypt formation and juvenile polyposis on BMP inhibition in mouse intestine. Science 303: 1684-1686.
40. Roth SI, Helwig EB (1963) Juvenile polyps of the colon and rectum. Cancer 16: 468-479.
41. Fearon ER, Vogelstein B (1990) A genetic model for colorectal tumorigenesis. Cell 61: 759-767.
42. Pino MS, Chung DC (2010) The chromosomal instability pathway in colon cancer. Gastroenterology 138: 2059-2072.
43. Poulogiannis G, Frayling IM, Arends MJ (2010) DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. Histopathology 56: 167-179.
44. (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330-337.
45. Lengauer C, Kinzler KW, Vogelstein B (1998) Genetic instabilities in human cancers. Nature 396: 643-649.
46. Ewing I, Hurley JJ, Josephides E, Millar A (2014) The molecular genetics of colorectal cancer. Frontline Gastroenterol 5: 26-30.
47. Bogaert J, Prenen H (2014) Molecular genetics of colorectal cancer. Ann Gastroenterol 27: 9-14.
48. Heinimann K (2013) Toward a molecular classification of colorectal cancer: the role of microsatellite instability status. Front Oncol 3: 272.
49. Issa JP (2004) CpG island methylator phenotype in cancer. Nat Rev Cancer 4: 988-993.
50. Nazemalhosseini Mojarad E, Kuppen PJ, Aghdaei HA, Zali MR (2013) The CpG island methylator phenotype (CIMP) in colorectal cancer. Gastroenterol Hepatol Bed Bench 6: 120-128.
51. Bettington M, Walker N, Clouston A, Brown I, Leggett B, et al. (2013) The serrated pathway to colorectal carcinoma: current concepts and challenges. Histopathology 62: 367-386.
52. Leggett B, Whitehall V (2010) Role of the serrated pathway in colorectal cancer pathogenesis. Gastroenterology 138: 2088-2100.
53. Jass JR (2007) Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 50: 113-130.
54. Houlston RS, Cheadle J, Dobbins SE, Tenesa A, Jones AM, et al. (2010) Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet 42: 973-977.
55. Schlussel AT, Gagliano RA, Jr., Seto-Donlon S, Eggerding F, Donlon T, et al. (2014) The evolution of colorectal cancer genetics-Part 1: from discovery to practice. J Gastrointest Oncol 5: 326-335.
56. Patel SG, Ahnen DJ (2012) Familial colon cancer syndromes: an update of a rapidly evolving field. Curr Gastroenterol Rep 14: 428-438.
57. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343: 78-85.
58. Trowbridge B, Burt RW (2002) Colorectal cancer screening. Surg Clin North Am 82: 943-957.
59. Fearnhead NS, Britton MP, Bodmer WF (2001) The ABC of APC. Hum Mol Genet 10: 721-733.
60. Jasperson KW, Tuohy TM, Neklason DW, Burt RW (2010) Hereditary and familial colon cancer. Gastroenterology 138: 2044-2058.

61. Half E, Bercovich D, Rozen P (2009) Familial adenomatous polyposis. Orphanet J Rare Dis 4: 22.
62. Knudsen AL, Bisgaard ML, Bulow S (2003) Attenuated familial adenomatous polyposis (AFAP). A review of the literature. Fam Cancer 2: 43-55.
63. Burt RW, Leppert MF, Slattery ML, Samowitz WS, Spirio LN, et al. (2004) Genetic testing and phenotype in a large kindred with attenuated familial adenomatous polyposis. Gastroenterology 127: 444-451.
64. Horii A, Nakatsuru S, Ichii S, Nagase H, Nakamura Y (1993) Multiple forms of the APC gene transcripts and their tissue-specific expression. Hum Mol Genet 2: 283-287.
65. Lambertz S, Ballhausen WG (1993) Identification of an alternative 5' untranslated region of the adenomatous polyposis coli gene. Hum Genet 90: 650-652.
66. De Rosa M, Morelli G, Cesaro E, Duraturo F, Turano M, et al. (2007) Alternative splicing and nonsense-mediated mRNA decay in the regulation of a new adenomatous polyposis coli transcript. Gene 395: 8-14.
67. Hosoya K, Yamashita S, Ando T, Nakajima T, Itoh F, et al. (2009) Adenomatous polyposis coli 1A is likely to be methylated as a passenger in human gastric carcinogenesis. Cancer Lett 285: 182-189.
68. Korinek V, Barker N, Morin PJ, van Wichen D, de Weger R, et al. (1997) Constitutive transcriptional activation by a beta-catenin-Tcf complex in APC-/- colon carcinoma. Science 275: 1784-1787.
69. Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, et al. (1997) Activation of beta-catenin-Tcf signaling in colon cancer by mutations in beta-catenin or APC. Science 275: 1787-1790.
70. Anastas JN, Moon RT (2013) WNT signalling pathways as therapeutic targets in cancer. Nat Rev Cancer 13: 11-26.
71. McNeill H, Woodgett JR (2010) When pathways collide: collaboration and connivance among signalling proteins in development. Nat Rev Mol Cell Biol 11: 404-413.
72. Roberts DM, Pronobis MI, Poulton JS, Waldmann JD, Stephenson EM, et al. (2011) Deconstructing the sscatenin destruction complex: mechanistic roles for the tumor suppressor APC in regulating Wnt signaling. Mol Biol Cell 22: 1845-1863.
73. Albuquerque C, Breukel C, van der Luijt R, Fidalgo P, Lage P, et al. (2002) The 'just-right' signaling model: APC somatic mutations are selected based on a specific level of activation of the beta-catenin signaling cascade. Hum Mol Genet 11: 1549-1560.
74. Lamlum H, Ilyas M, Rowan A, Clark S, Johnson V, et al. (1999) The type of somatic mutation at APC in familial adenomatous polyposis is determined by the site of the germline mutation: a new facet to Knudson's 'two-hit' hypothesis. Nat Med 5: 1071-1075.
75. Latchford A, Volikos E, Johnson V, Rogers P, Suraweera N, et al. (2007) APC mutations in FAP-associated desmoid tumours are non-random but not 'just right'. Hum Mol Genet 16: 78-82.
76. Segditsas S, Tomlinson I (2006) Colorectal cancer and genetic alterations in the Wnt pathway. Oncogene 25: 7531-7537.
77. Heinen CD (2010) Genotype to phenotype: analyzing the effects of inherited mutations in colorectal cancer families. Mutat Res 693: 32-45.
78. Morak M, Heidenreich B, Keller G, Hampel H, Laner A, et al. (2014) Biallelic MUTYH mutations can mimic Lynch syndrome. Eur J Hum Genet.

79. Venesio T, Balsamo A, D'Agostino VG, Ranzani GN (2012) MUTYH-associated polyposis (MAP), the syndrome implicating base excision repair in inherited predisposition to colorectal tumors. Front Oncol 2: 83.
80. Win AK, Dowty JG, Cleary SP, Kim H, Buchanan DD, et al. (2014) Risk of colorectal cancer for carriers of mutations in MUTYH, with and without a family history of cancer. Gastroenterology 146: 1208-1211.e1201-1205.
81. David SS, O'Shea VL, Kundu S (2007) Base-excision repair of oxidative DNA damage. Nature 447: 941-950.
82. Nakabeppu Y, Tsuchimoto D, Furuichi M, Sakumi K (2004) The defense mechanisms in mammalian cells against oxidative damage in nucleic acids and their involvement in the suppression of mutagenesis and cell death. Free Radic Res 38: 423-429.
83. Jelsig AM, Qvist N, Brusgaard K, Nielsen CB, Hansen TP, et al. (2014) Hamartomatous polyposis syndromes: a review. Orphanet J Rare Dis 9: 101.
84. Gammon A, Jasperson K, Kohlmann W, Burt RW (2009) Hamartomatous polyposis syndromes. Best Pract Res Clin Gastroenterol 23: 219-231.
85. Jenne DE, Reimann H, Nezu J, Friedel W, Loff S, et al. (1998) Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. Nat Genet 18: 38-43.
86. Brosens LA, van Hattem A, Hylind LM, Iacobuzio-Donahue C, Romans KE, et al. (2007) Risk of colorectal cancer in juvenile polyposis. Gut 56: 965-967.
87. Jass JR (2007) Gastrointestinal polyposes: clinical, pathological and molecular features. Gastroenterol Clin North Am 36: 927-946, viii.
88. Howe JR, Bair JL, Sayed MG, Anderson ME, Mitros FA, et al. (2001) Germline mutations of the gene encoding bone morphogenetic protein receptor 1A in juvenile polyposis. Nat Genet 28: 184-187.
89. Howe JR, Roth S, Ringold JC, Summers RW, Jarvinen HJ, et al. (1998) Mutations in the SMAD4/DPC4 gene in juvenile polyposis. Science 280: 1086-1088.
90. Latchford AR, Neale K, Phillips RK, Clark SK (2012) Juvenile polyposis syndrome: a study of genotype, phenotype, and long-term outcome. Dis Colon Rectum 55: 1038-1043.
91. Ngeow J, Heald B, Rybicki LA, Orloff MS, Chen JL, et al. (2013) Prevalence of germline PTEN, BMPR1A, SMAD4, STK11, and ENG mutations in patients with moderate-load colorectal polyps. Gastroenterology 144: 1402-1409, 1409.e1401-1405.
92. Lucci-Cordisco E, Risio M, Venesio T, Genuardi M (2013) The growing complexity of the intestinal polyposis syndromes. Am J Med Genet A 161a: 2777-2787.
93. Pilarski R (2009) Cowden syndrome: a critical review of the clinical literature. J Genet Couns 18: 13-27.
94. Jaeger E, Leedham S, Lewis A, Segditsas S, Becker M, et al. (2012) 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 44: 699-703.
95. Cheah PY, Wong YH, Chau YP, Loi C, Lim KH, et al. (2009) Germline bone morphogenesis protein receptor 1A mutation causes colorectal tumorigenesis in hereditary mixed polyposis syndrome. Am J Gastroenterol 104: 3027-3033.
96. Gala MK, Mizukami Y, Le LP, Moriichi K, Austin T, et al. (2014) Germline mutations in oncogene-induced senescence pathways are associated with multiple sessile serrated adenomas. Gastroenterology 146: 520-529.

97. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, et al. (2013) Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat Genet 45: 136-144.
98. Briggs S, Tomlinson I (2013) Germline and somatic polymerase epsilon and delta mutations define a new class of hypermutated colorectal and endometrial cancers. J Pathol 230: 148-153.
99. Rohlin A, Zagoras T, Nilsson S, Lundstam U, Wahlstrom J, et al. (2014) A mutation in POLE predisposing to a multi-tumour phenotype. Int J Oncol 45: 77-81.
100. Heitzer E, Tomlinson I (2014) Replicative DNA polymerase mutations in cancer. Curr Opin Genet Dev 24: 107-113.
101. Dunlop MG, Dobbins SE, Farrington SM, Jones AM, Palles C, et al. (2012) Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat Genet 44: 770-776.
102. Church DN, Briggs SE, Palles C, Domingo E, Kearsey SJ, et al. (2013) DNA polymerase epsilon and delta exonuclease domain mutations in endometrial cancer. Hum Mol Genet 22: 2820-2828.
103. Seshagiri S (2013) The burden of faulty proofreading in colon cancer. Nat Genet 45: 121-122.
104. Shiovitz S, Copeland WK, Passarelli MN, Burnett-Hartman AN, Grady WM, et al. (2014) Characterisation of familial colorectal cancer Type X, Lynch syndrome, and non-familial colorectal cancer. Br J Cancer 111: 598-602.
105. Lindor NM (2009) Familial colorectal cancer type X: the other half of hereditary nonpolyposis colon cancer syndrome. Surg Oncol Clin N Am 18: 637-645.
106. Hitchins MP (2013) The role of epigenetics in Lynch syndrome. Fam Cancer 12: 189-205.
107. Peltomaki P (2014) Epigenetic mechanisms in the pathogenesis of Lynch syndrome. Clin Genet 85: 403-412.
108. Morak M, Schackert HK, Rahner N, Betz B, Ebert M, et al. (2008) Further evidence for heritability of an epimutation in one of 12 cases with MLH1 promoter methylation in blood cells clinically displaying HNPCC. Eur J Hum Genet 16: 804-811.
109. Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, et al. (2009) Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. Nat Genet 41: 112-117.
110. Niessen RC, Hofstra RM, Westers H, Ligtenberg MJ, Kooi K, et al. (2009) Germline hypermethylation of MLH1 and EPCAM deletions are a frequent cause of Lynch syndrome. Genes Chromosomes Cancer 48: 737-744.
111. Capper D, Voigt A, Bozukova G, Ahadova A, Kickingereder P, et al. (2013) BRAF V600E-specific immunohistochemistry for the exclusion of Lynch syndrome in MSI-H colorectal cancer. Int J Cancer 133: 1624-1630.
112. Lynch HT, Lynch PM, Lanspa SJ, Snyder CL, Lynch JF, et al. (2009) Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin Genet 76: 1-18.
113. Niittymaki I, Kaasinen E, Tuupanen S, Karhu A, Jarvinen H, et al. (2010) Low-penetrance susceptibility variants in familial colorectal cancer. Cancer Epidemiol Biomarkers Prev 19: 1478-1483.
114. Lewis A, Freeman-Mills L, de la Calle-Mustienes E, Giraldez-Perez RM, Davis H, et al. (2014) A polymorphic enhancer near GREM1 influences bowel cancer risk through differential CDX2 and TCF7L2 binding. Cell Rep 8: 983-990.

115. Tenesa A, Dunlop MG (2009) New insights into the aetiology of colorectal cancer from genome-wide association studies. Nat Rev Genet 10: 353-358.
116. Rennert G, Almog R, Tomsho LP, Low M, Pinchev M, et al. (2005) Colorectal polyps in carriers of the APC I1307K polymorphism. Dis Colon Rectum 48: 2317-2321.
117. Hazra A, Fuchs CS, Chan AT, Giovannucci EL, Hunter DJ (2008) Association of the TCF7L2 polymorphism with colorectal cancer and adenoma risk. Cancer Causes Control 19: 975-980.
118. Xu Y, Pasche B (2007) TGF-beta signaling alterations and susceptibility to colorectal cancer. Hum Mol Genet 16 Spec No 1: R14-20.
119. Clarke E, Green RC, Green JS, Mahoney K, Parfrey PS, et al. (2012) Inherited deleterious variants in GALNT12 are associated with CRC susceptibility. Hum Mutat 33: 1056-1058.
120. Nieminen TT, O'Donohue MF, Wu Y, Lohi H, Scherer SW, et al. (2014) Germline Mutation of RPS20, Encoding a Ribosomal Protein, Causes Predisposition to Hereditary Nonpolyposis Colorectal Carcinoma Without DNA Mismatch Repair Deficiency. Gastroenterology 147: 595-598.e595.
121. Schulz E, Klampfl P, Holzapfel S, Janecke AR, Ulz P, et al. (2014) Germline variants in the SEMA4A gene predispose to familial colorectal cancer type X. Nat Commun 5: 5191.
122. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, et al. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51 Pt 1: 263-273.
123. Vogelstein B, Kinzler KW (1999) Digital PCR. Proc Natl Acad Sci U S A 96: 9236-9241.
124. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74: 5463-5467.
125. Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11: 31-46.
126. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.
127. Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9: 387-402.
128. Liu L, Li Y, Li S, Hu N, He Y, et al. (2012) Comparison of next-generation sequencing systems. J Biomed Biotechnol 2012: 251364.
129. Dolan PC, Denver DR (2008) TileQC: a system for tile-based quality control of Solexa data. BMC Bioinformatics 9: 250.
130. Schroder J, Bailey J, Conway T, Zobel J (2010) Reference-free validation of short read data. PLoS One 5: e12681.
131. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, et al. (2011) Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res 39: e90.
132. Aird D, Ross MG, Chen WS, Danielsson M, Fennell T, et al. (2011) Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol 12: R18.
133. Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP (2014) Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet 15: 121-132.
134. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.

135. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
136. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43: 491-498.
137. Zhao M, Wang Q, Wang Q, Jia P, Zhao Z (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics 14 Suppl 11: S1.
138. Tan R, Wang Y, Kleinstein SE, Liu Y, Zhu X, et al. (2014) An evaluation of copy number variation detection tools from whole-exome sequencing data. Hum Mutat 35: 899-907.
139. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40: e72.
140. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38: e164.
141. Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, et al. (2002) Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res 30: e57.
142. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277-318.
143. Galiatsatos P, Foulkes WD (2006) Familial adenomatous polyposis. Am J Gastroenterol 101: 385-398.
144. Jarvinen HJ, Peltomaki P (2004) The complex genotype-phenotype relationship in familial adenomatous polyposis. Eur J Gastroenterol Hepatol 16: 5-8.
145. Aretz S, Uhlhaas S, Caspari R, Mangold E, Pagenstecher C, et al. (2004) Frequency and parental origin of de novo APC mutations in familial adenomatous polyposis. Eur J Hum Genet 12: 52-58.
146. Gayther SA, Wells D, SenGupta SB, Chapman P, Neale K, et al. (1994) Regionally clustered APC mutations are associated with a severe phenotype and occur at a high frequency in new mutation cases of adenomatous polyposis coli. Hum Mol Genet 3: 53-56.
147. Renkonen ET, Nieminen P, Abdel-Rahman WM, Moisio AL, Jarvela I, et al. (2005) Adenomatous polyposis families that screen APC mutation-negative by conventional methods are genetically heterogeneous. J Clin Oncol 23: 5651-5659.
148. Pavicic W, Nieminen TT, Gylling A, Pursiheimo JP, Laiho A, et al. (2014) Promoter-specific alterations of APC are a rare cause for mutation-negative familial adenomatous polyposis. Genes Chromosomes Cancer 53: 857-864.
149. Snow AK, Tuohy TM, Sargent NR, Smith LJ, Burt RW, et al. (2014) APC promoter 1B deletion in seven American families with familial adenomatous polyposis. Clin Genet.
150. Kadiyska TK, Todorov TP, Bichev SN, Vazharova RV, Nossikoff AV, et al. (2014) APC promoter 1B deletion in familial polyposis--implications for mutation-negative families. Clin Genet 85: 452-457.
151. Charames GS, Ramyar L, Mitri A, Berk T, Cheng H, et al. (2008) A large novel deletion in the APC promoter region causes gene silencing and leads to classical familial adenomatous polyposis in a Manitoba Mennonite kindred. Hum Genet 124: 535-541.

152. Necker J, Kovac M, Attenhofer M, Reichlin B, Heinimann K (2011) Detection of APC germ line mosaicism in patients with de novo familial adenomatous polyposis: a plea for the protein truncation test. J Med Genet 48: 526-529.
153. Coppin L, Grutzmacher C, Crepin M, Destailleur E, Giraud S, et al. (2014) VHL mosaicism can be detected by clinical next-generation sequencing and is not restricted to patients with a mild phenotype. Eur J Hum Genet 22: 1149-1152.
154. Pritchard CC, Smith C, Marushchak T, Koehler K, Holmes H, et al. (2013) A mosaic PTEN mutation causing Cowden syndrome identified by deep sequencing. Genet Med 15: 1004-1007.
155. Izawa K, Hijikata A, Tanaka N, Kawai T, Saito MK, et al. (2012) Detection of base substitution-type somatic mosaicism of the NLRP3 gene with >99.9% statistical confidence by massively parallel sequencing. DNA Res 19: 143-152.
156. Potter JD (1999) Colorectal cancer: molecules and populations. J Natl Cancer Inst 91: 916-932.
157. Kosinski C, Li VS, Chan AS, Zhang J, Ho C, et al. (2007) Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proc Natl Acad Sci U S A 104: 15418-15423.
158. Hardwick JC, Kodach LL, Offerhaus GJ, van den Brink GR (2008) Bone morphogenetic protein signalling in colorectal cancer. Nat Rev Cancer 8: 806-812.
159. Scoville DH, Sato T, He XC, Li L (2008) Current view: intestinal stem cells and signaling. Gastroenterology 134: 849-864.
160. Valle L, Hernandez-Illan E, Bellido F, Aiza G, Castillejo A, et al. (2014) New insights into POLE and POLD1 germline mutations in familial colorectal cancer and polyposis. Hum Mol Genet 23: 3506-3512.
161. Elsayed FA, Kets CM, Ruano D, van den Akker B, Mensenkamp AR, et al. (2014) Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer.
162. Cragun D, Radford C, Dolinsky JS, Caldwell M, Chao E, et al. (2014) Panel-based testing for inherited colorectal cancer: a descriptive study of clinical testing performed by a US laboratory. Clin Genet.
163. Kraus C, Rau TT, Lux P, Erlenbach-Wunsch K, Lohr S, et al. (2014) Comprehensive screening for mutations associated with colorectal cancer in unselected cases reveals penetrant and nonpenetrant mutations. Int J Cancer.
164. LaDuca H, Stuenkel AJ, Dolinsky JS, Keiles S, Tandy S, et al. (2014) Utilization of multigene panels in hereditary cancer predisposition testing: analysis of more than 2,000 patients. Genet Med 16: 830-837.