Characteristics of the germ-line Copy Number Variations (CNVs) in colorectal cancer patients

By

© Salem Werdyani

A thesis submitted to School of Graduate Studies

in partial fulfillment of the requirement for the degree of

Master of Science in Medicine (Human Genetics)

Discipline of Genetics/Faculty of Medicine

Memorial University of Newfoundland

May 2016

St. John's Newfoundland

ABSTRACT

INDELs and CNVs are types of structural variations. This project aimed to computationally predict and examine the possible biological consequences of the INDELs/CNVs in colorectal cancer patient genomes using the QuantiSNP and PennCNV algorithms. A total of 69,290 INDELs/CNVs were identified in 495 patients and satisfied the quality control criteria. These variations constituted 3,486 distinct INDELs/CNVs and clustered in 1,527 CNV regions. The majority of the variations were CNVs (~79%) and deletions (~81%). Around 63% of the distinct variations were identified to completely or partially cover the sequences of 1,673 genes, and a large number of these genes were observed to act in cancer related biological pathways. In summary, this project detected a number of biologically interesting INDELs/CNVs in the genomes of colorectal cancer patients. Further studies in the Savas lab will investigate these variants in detail including their possible associations with the disease characteristics and outcomes in colorectal cancer.

ACKNOWLEDGEMENT

In the name of Allah, the Most Compassionate and the Most Merciful. All praises, gratitude and glory to Almighty Allah who gave me the strengths, courage and blessing throughout my research work to complete this thesis successfully, and peace and blessing of Allah be upon our beloved Prophet Mohammad (S).

I would like to express my heartfelt gratitude and unrestrained appreciation to my supervisor, Dr. Sevtap Savas, whose encouragement, supervision, guidance and unlimited support throughout my thesis project have led to the success of this research. I also gratefully acknowledge the financial support of the School of Graduate Studies, Research and Graduate Studies at the Faculty of Medicine during my Master's program. Moreover, special appreciation goes to Colon Cancer Canada that funded this thesis project.

My sincere thanks to those people who in one way or another contributed and extended their support and valuable assistance in the preparation and completion of this academic work, including my supervisory committee members (the late Dr. Roger Green, Dr. Guangju Zhai, and Dr. Michael Woods), and the members of the Newfoundland Colorectal Cancer Registry (NFCCR). Furthermore, I wish to convey my thanks to Dr. Kai Wang (University of Pennsylvania) for promptly responding to my queries regarding the CNV prediction using the PennCNV algorithm.

My gratitude and thanks are extended to my Beloved parents, Adnan and Lamah Werdyani for their infinite sacrifice, encouragement and prayers; my parents in law, Yehia Alkhateb and Nahed Haj Ali for their support and blessings; and all members of my family. I would like also to take this opportunity to express the grateful appreciation to my wife, Dr. Rahaf Alkhateb who never failed to provide me with love, warmth, patience and care during my thesis project. Not forgotten, a special thanks to my dearest children, Adnan and Mohamed Nour Werdyani, who have been my inspiration as I overcome all the obstacles in the completion of this thesis project.

Salem Werdyani.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABBREVIATIONS

A

aCGH	Array comparative genomic hybridization
adh	Short-chain dehydrogenase
AIDS	Acquired immunodeficiency syndrome
AFAP	Attenuated familial adenomatous polyposis
APC	Adenomatous polyposis coli
ATM	Ataxia telangiectasia mutated

B

BAC	Bacterial artificial chromosome
BAF	B allele frequency
BAF_SD	B allele frequency standard deviation
BAF_Drift	B allele frequency drift
BF	Bayes factor
BIR	Break-induced replication
BMI	Body mass index
BMP2	Bone morphogenetic protein 2
BMP4	Bone morphogenetic protein 4
BMPR1A	Bone morphogenetic protein receptor type 1A
bp	Base pair

C

CCL3L1	Human CC chemokine ligand 3-like 1
CDH1	Cadherin 1
CEU	Utah Residents with Northern and Western European ancestry (Caucasian population)
CFHR2	Complement Factor H-Related 2
CGH	Comparative genomic hybridization
CHB	Han Chinese from Beijing, China

CHEK2	Checkpoint Kinase 2
CIHR	The Canadian Institutes of Health Research
CNP	Copy number polymorphism
CN	Copy number
CNV	Copy number variation
CNVR	Copy number variation region
CS	Cowden's syndrome
CYP2A7	Cytochrome P450, Family 2, Subfamily A, Polypeptide 7
c8orf53	Small subunit processome component homolog
c11orf53	Chromosome 11 open reading frame 53

D

ddNTP	Dideoxynucleotide triphosphates
DHJ	Double Holliday junction
DGV	Database of Genomic Variants
DLEC1	Deleted in lung and esophageal cancer 1
DNA	Deoxyribonucleic acid
DSB	Double-stand break

E

EGF	Epidermal growth factor
EGFR	Epidermal growth factor receptor
EIF3H	Eukaryotic translation initiation factor 3 subunit H
EM	Expectation maximization
ENCODE	Encyclopedia of DNA Elements
ERBB2	Receptor tyrosine-protein kinase erbB-2
ERBB3	Receptor tyrosine-protein kinase erbB-3

F

FAP	Familial adenomatous polyposis

FCCX	Familial colorectal cancer type X
FISH	Fluorescent in situ hybridization
FoSTeS	Fork stalling and template switching
FLJ45803	Colorectal cancer associated 1
FLT3	Fms-like tyrosine kinase 3
FMR1	Fragile X Mental Retardation 1
FSTL5	Follistatin-Like 5 gene

G

|GCWF|	The absolute GC waviness factor
GEO	Gene Expression Omnibus database
GRCh36	Genome Reference Consortium build 36
GRCh37	Genome Reference Consortium build 37
GREM1	Germlin 1, DNA family BMP antagonist
GWAS	Genome wide association study

H

HEATR4	HEAT induced repeat containing
Hg19	Human genome assembly hg19
Hg18	Human genome assembly hg18
HIV	Human immunodeficiency virus
HLA-A	Major histocompatibility complex, class I, A
HMM	Hidden Markov model
HNPCC	Hereditary nonpolyposis colorectal cancer
HORDE	The Human Olfactory Receptors Data Explorer database
HR	Homologous recombination
HREA	Health Research Ethics Authority

I

IBD	Inflammatory bowel disease

INDEL	Insertion/deletion
IGF2	Insulin-like growth factor 2
IFN	Interferon signaling pathway
IFNAR1	Interferon (alpha, beta and omega) receptor 1
IFNGR1	Interferon gamma receptor 1
IFNA14	Interferon alpha 14

J

JPS	Juvenile polyposis syndrome
JPT	Japanese from Tokyo, Japan

K

kbp	Kilobase pair
KIF26B	Kinesin Family Member 26B
KRAS	Kirsten rat sarcoma viral oncogene homolog

L

lincRNA	Long, intergenic non-coding RNA
LOC120376	Colorectal cancer associated 2
LOH	Loss of heterozygosity
LRR	Log_2 R ratio
LRR_SD	Log R ratio standard deviation

M

MAF	Minor allele frequency
Mbp	Megabase pair
miRNA	microRNA precursors
miscRNA	miscellaneous other RNA
MLH1	MutL homolog 1, mismatch repair
MLH3	MutL homolog 3, mismatch repair

MMBIR	Microhomology-mediated break-induced replication
MMEJ	Microhomology-mediated end joining
MMR	Mismatch repair
MSH2	MutS homolog 2, mismatch repair
MSH6	MutS homolog 6, mismatch repair
MSI	Microsatellite instability
MSS	Microsatellite stable
MTS	Muir-Torre syndromes
Mtus1	Mitochondrial tumor suppressor
MUTYH	MutY DNA glycosylase
MYC	V-myc avian myelocytomatosis viral oncogene homolog

N

NAHR	Non-allelic homologous recombination
NCBI	The National Center for Biotechnology Information
NFCCR	Newfoundland Colorectal Cancer Registry
NFKB	Nuclear factor of kappa B
NHEJ	Nonhomologous end joining
NHR	Nonhomologous recombination
NL	Newfoundland and Labrador
NME7	Nonmetastatic cells, 7
NPM1	Nucleophosmin 1

O

OB-HMM	Objective Bayes Hidden-Markov model
OR	Olfactory receptor

P

PCDHA9	Protocadhein Alpha 9
PCR	Polymerase chain reaction

PFB	Population frequency of the B allele
PJS	Peutz-Jeghers syndrome
PMS1	Postmeiotic segregation increased 1, mismatch repair
PMS2	Postmeiotic segregation increased 2, mismatch repair
POU2AF1	POU class 2 associating factor 1
P53	P53 tumor suppressor protein

Q

QC	Quality control

R

R	The total normalized intensity value of A and B alleles
RHPN2	Rho GTPase binding protein 2
RPS20	Ribosomal protein S20
rRNA	ribosomal RNAs
RSPO2	R-spondin 2
RSPO3	R-spondin 3

S

SD	Standard deviation
SegDup	Segmental duplication
SIRPB1	Signal-regulatory protein beta 1
SMAD4	Mothers against decapentaplegic homolog 4
SMAD7	Mothers against decapentaplegic homolog 7
snRNA	small nuclear RNAs
snoRNA	small nucleolar RNAs
SNP	Single nucleotide polymorphism
STR	Short tandem repeat
SSA	Single stand annealing
SV	Structural variation

T

TCF7L2	Transcription factor 7-Like 2
TET2	Tet methylcytosine dioxygenase 2
TET3	Tet methylcytosine dioxygenase 3
TGF-β	Transforming Growth Factor β
Theta (θ)	The allelic intensity ratio
TSG	Tumor suppressor gene
TUBA8	Tubulin, Alpha 8
γTuRC	γ-tubulin ring complex

U

UCSC	University of California, Santa Cruz

V

VEGF	Vascular endothelial growth factor

W

WF	Waviness factor
WWOX	WW domain containing oxidoreductase

X

Y

YRI	Yoruba from Ibadan population, Nigeria

Z

RESEARCH OUTPUT AND AWARDS

Abstracts

A. Oral presentations

1. **<u>Salem Werdyani</u>**, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Patrick Parfrey, Roger Green, Sevtap Savas. *Features of the Copy Number Variants* (*CNVs) in a cohort of colorectal cancer patients*. The 5th UAE National Genetic Diseases Conference, September 13-17, 2014, Dubai, United Arab Emirates (Invited keynote speaker).
2. **<u>Salem Werdyani</u>**, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Patrick Parfrey, Roger Green, Sevtap Savas. *Genes and pathways that may be affected by the germline Copy Number Variations (CNVs) in colorectal cancer patients*. The 3rd Annual Canadian Human and Statistical Genetics Meeting, May 3-6, 2014, Victoria, BC, Canada (Selected for oral presentation).

B. Poster presentations

1. **<u>Salem Werdyani</u>**, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Patrick Parfrey, Roger Green, Sevtap Savas. *Copy Number Variants (CNVs) and small Insertions/Deletions (INDELs) in a cohort of colorectal cancer patients from Newfoundland.* The 4th Annual Canadian Human and Statistical Genetics Meeting, April 18-21, 2015, Vancouver, BC, Canada.
2. **<u>Salem Werdyani</u>**, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Patrick Parfrey, Roger Green, Sevtap Savas. *Copy*

Number Variations (CNVs) and colorectal cancer. The 2015 Canadian Cancer Research Conference, November 8-10, 2015, Montreal, QC, Canada.

C. Other presentations as a co-author

1. Yajun Yu, **Salem Werdyani**, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Yildiz Yilmaz, Patrick Parfrey, Sevtap Savas. *Structural variants in TGFBR3, STEAP2, and FILIP1L genes may associate with disease outcomes in colorectal cancer.* The Target Meeting's 4th World Cancer Online Conference, May 17-19, 2016 (submitted).
2. Yajun Yu, **Salem Werdyani**, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Yildiz Yilmaz, Patrick Parfrey, Sevtap Savas. *Common Copy Number Variations (CNVs) and disease-free survival in colorectal cancer.* The 5th Annual Canadian Human and Statistical Genetics Meeting, April 16-19, 2016, Halifax, NS, Canada (submitted).
3. Sevtap Savas, Jingxiong Xu, **Salem Werdyani**, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu. *Replication of associations of two polymorphisms with survival times in colorectal cancer.* The 2nd International Conference on Predictive, Preventive and Personalized Medicine & Molecular Diagnostics, Nov 3-5, 2014, Las Vegas, USA (invited speech by Dr. Sevtap Savas).
4. Lydia Dan, Jingxiong Xu, **Salem Werdyani**, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu, Sevtap Savas. *Genetic polymorphisms in matrix metalloproteinase genes MMP8 and MMP27 are associated with overall survival in colorectal cancer*. The Target Meeting's 3rd

World Cancer Online Conference, January 21-24, 2014 (oral presentation by Ms. Lydia Dan).

5. Sevtap Savas, Lydia Dan, Jingxiong Xu, **Salem Werdyani**, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu. *Genetic polymorphisms and outcome research in cancer: Examples from angiogenesis and metastasis genes and colorectal cancer*. The 2nd International Conference on Predictive, Preventive and Personalized Medicine & Molecular Diagnostics, Nov 3-5, 2014, Las Vegas, USA (poster presentation by Dr. Sevtap Savas).
6. Lydia Dan, Jingxiong Xu, **Salem Werdyani**, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu, Sevtap Savas. *Prognostic association of the polymorphisms in matrix metalloproteinase genes MMP8 and MMP27 in patients with colorectal cancer*. The 3rd Annual Canadian Human and Statistical Genetics Meeting, May 3-6, 2014, Victoria, BC, Canada (poster presentation by Ms. Lydia Dan).
7. Lydia Dan, Jingxiong Xu, **Salem Werdyani**, Konstantin Shestopaloff, Elizabeth Dicks, Patrick Parfrey, Roger Green, Wei Xu, Sevtap Savas. *Genetic polymorphisms in angiogenesis, lymph-angiogenesis, and metastasis pathway genes and the disease outcome in colorectal cancer*. Poster presentation in the 2013 Canadian Cancer Research Conference, November 2-6, 2013, Toronto, ON, Canada (poster presentation by Ms. Lydia Dan).

Publications

A. Published

1. Sevtap Savas, Jingxiong Xu, **Salem Werdyani**, Konstantin Shestopaloff, Elizabeth Dicks, Jane Green, Patrick Parfrey, Roger Green, Wei Xu. *A survival association study of 102 polymorphisms previously associated with survival outcomes in colorectal cancer.* BioMed Research International, Volume 2015 (2015), Article ID 968743.

B. In preparation

1. **Salem Werdyani**, Georgia Skardasi, Jingxiong Xu, Konstantin Shestopaloff, Wei Xu, Elizabeth Dicks, Jane Green, Patrick Parfrey, Roger Green, Sevtap Savas. *Copy Number Variations (CNVs) and colorectal cancer* (In preparation).
2. Lydia Dan, **Salem Werdyani**, Jingxiong Xu, Konstantin Shestopaloff, Elizabeth Dicks, Jane Green, Patrick Parfrey, Roger Green, Wei Xu, Sevtap Savas. *Genetic polymorphisms in angiogenesis, lymph-angiogenesis and metastasis pathway genes and the disease outcome in colorectal cancer* (ready to be submitted).

Travel Awards

1. May 2014: I was awarded the first inaugural MGSS Graduate Travel Award in Medicine, Faculty of Medicine, Memorial University of Newfoundland ($500).
2. July 2015: I was awarded a Travel Award from - the Institute Community Support program of the Canadian Institute of Health Research ($1,000).

Travel Support

1. April 2015: Travel support provided by the Discipline of Oncology is gratefully acknowledged ($1,650).
2. July 2015: Travel bursary provided by the Terry Fox Research Institute (TFRI) and Beatrice Hunter Cancer Research Institute (BHCRI) ($500).

Chapter 1

1.1 Overview of the research project

Colorectal cancer is the abnormal growth of the epithelial cells in the colon or rectum. It is estimated to be one of the major malignancies worldwide and the second leading cause of cancer-related death for both sexes in Canada [1-3]. Among the Canadian provinces, Newfoundland and Labrador (NL) has the highest incidence and mortality rates of this disease [3].

Copy Number Variations (CNVs) are recently discovered genetic variations that consist of large deletions, insertions, or duplications of DNA fragments existing in different copy numbers among individuals. CNVs occur frequently in the human genome and may affect the expression and function of genes. Several studies suggested a significant contribution of CNVs in human phenotypic variability including in susceptibility to diseases, such as autism [4], Alzheimer's and Parkinson's diseases [5,6], and cancer [7].

CNVs have been identified to affect 4-15% of the cancer-related genes [8]. For example, a 3,670 base pairs (bps) germ-line deletion on 2p24.3 was reported to be significantly associated with the risk of developing aggressive prostate cancer [9]. Another deletion CNV at 22q12.1 removes the exons 9 and 10 of the checkpoint Kinase 2 (*CHEK2*) breast cancer suppressor gene, and was reported to double the risk of breast cancer in several populations [10,11]. Some studies have also reported the important roles of CNVs in the susceptibility and prognosis of colorectal cancer [12-14]. For instance, a heterozygous 4 kilobase pairs (kbps) germline deletion covering the exon 9 of a tumor-associated calcium signal transducer 1 (*TACSTD1*) gene has been identified to significantly affect the expression of the MutS Homolog 2 (*MSH2*) gene, which is the gene deficient in Lynch syndrome [14].

This thesis project constitutes the initial parts of a larger project that aims to investigate the CNVs in relation to colorectal cancer. The main objectives of this project were: a) to computationally predict and characterize the CNV profiles in the genomes of a cohort of 505 Caucasian colorectal cancer patients [15,16], and b) to identify the genes and biological pathways that may be affected by the detected variants. A previously generated genome-wide SNP genotyping and signal intensity data were used to detect CNVs in the genome of each patient using the QuantiSNP and PennCNV algorithms [17,18]. Although this approach aimed to detect CNVs, small insertion/deletion variants (INDELs) were also detected. To exclude low quality data and to reduce false-positive findings, a group of stringent quality control (QC) analyses was performed by using programming languages, such as Java and Perl. The data from the individuals and variations that fulfilled these QC analyses were then used to identify the human genes and biological pathways that were possibly affected by the INDELs/CNVs. These results identified a large number of biologically interesting INDELs/CNVs that will be significant for the worldwide scientific community.

1.2 Human genetic variations

The inheritance of traits has been an interesting research area. In the early 1900's, it became obvious that multiple quantitative traits and diseases can be transmitted within the same family throughout generations. In addition, it was observed that close relatives shared more trait similarities with each other than distant family members [19]. After the discovery of the chromosomal basis of inheritance, variations in DNA have been considered to be the cause of heritable phenotypes [20].

Although about 99.9% of the DNA sequences are identical between any two randomly selected humans, the remaining 0.1% of the genome underlies the genotypic diversity among individuals and populations [21]. Genetic variations can be inherited or newly formed (*de novo*). If a genetic variant is formed in gametes, it is called a germ-line variant, and it may be passed to the next generations [22]. In contrast, if a variant occurs in somatic tissues during development, it is called a somatic variation, which cannot be passed on to the next generations [23].

Human genetic variations range in size from single nucleotide changes and tandem repeats to small insertions/deletions (INDELs) and copy number variations (CNVs) of large DNA segments [24,25]. These variations exist in two or more alleles at a locus. Traditionally, variations are categorized as rare or common based on the frequency of their least frequent allele, which is called the minor allele frequency (MAF) [26]. Different studies have classified genetic variations based on different MAF thresholds [27]; however, throughout this thesis project, variants with a MAF $< 5\%$ are considered to be rare variants, whereas variants having a MAF $\geq 5\%$ are classified as common variants. Additionally, when the MAF of a genetic variant is $\geq 1\%$ in the population, it is called a polymorphism [28].

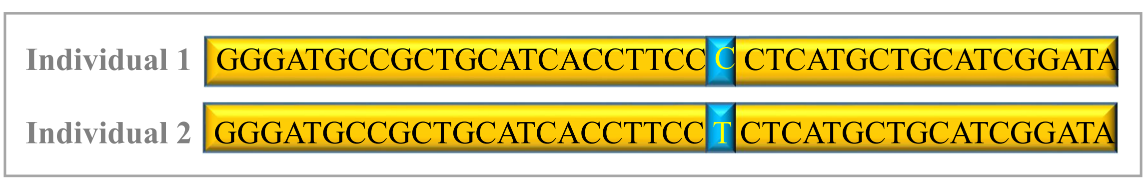
In the following section, some of the main types of genetic variations are discussed in detail.

1.2.1 Single nucleotide polymorphisms (SNPs)

SNPs are substitutions of a single nucleotide (**Figure 1.1**). SNPs are often bi-allelic, consisting of two alleles; A and B. These alleles can present in three different allele combinations or genotypes; homozygous genotypes AA or BB, and heterozygous genotype AB [29].

SNPs are the most common and stable genetic variations that are distributed throughout the human genome [30]. Nearly 15 million SNPs have been identified in human populations [31]. Up to 12% of the SNPs have been estimated to fall within coding sequences, while the rest are located in the noncoding DNA regions. SNPs in the coding and regulatory regions are more likely to affect the biological functions of the genes and contribute to the phenotypic variability [32]. The Encyclopedia of DNA Elements (ENCODE) project has estimated that 80% of human bases have biochemical roles in at least one tissue; these findings led to an increased interest in SNPs that are located in the noncoding regions [33,34].

Figure 1.1: Single nucleotide polymorphism (SNP).



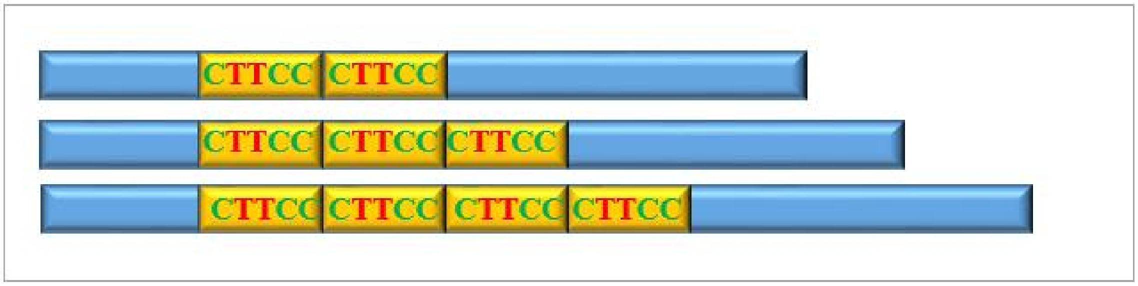
SNP is a single base change along the DNA sequence. The bases, C and T, in this example illustrate the different alleles at a SNP locus.

Because of their likely biological importance and their abundance in the human genome, together with the advanced technologies used to genotype their alleles [35], SNPs have served as excellent genetic markers in the genome-wide association studies (GWASs). In GWAS, SNPs are studied for their potential roles in complex genetic traits, disease susceptibility, and response to drugs [36,37]. A variety of human diseases, such as high blood pressure, asthma, and type 2 diabetes, have been identified to associate with SNPs [38-41]. Furthermore, multiple studies reported that SNPs can modify the risk of cancer [42-45]. For example, Broderick et al (2007) identified a highly significant association between the SNP rs4939827 in the mothers against decapentaplegic homolog 7 (*SMAD7*) gene and colorectal cancer [42]. Another example includes the SNP rs2856968 in the interferon (alpha, beta and omega) receptor 1 (*IFNAR1*) gene and the SNP rs2234711 in the interferon gamma receptor 1 (*IFNGR1*) gene, which are significantly associated with the elevated risk of colorectal cancer [43]. Additionally, the SNP rs6475526 located ~ two kbps upstream of the interferon alpha 14 (*IFNA14*) gene has been reported to significantly associate with the overall survival times of colorectal cancer patients [43].

1.2.2 Microsatellites

DNA variations that consist of short repeats of DNA sequences are called short tandem repeats (STR), or microsatellite repeats [46,47]. Microsatellites consist of one to six nucleotides that are tandemly repeated several times [48](**Figure 1.2)**. For example, $(CA)_n$ repeats are well known microsatellites [49].

Figure 1.2: Examples of microsatellites.



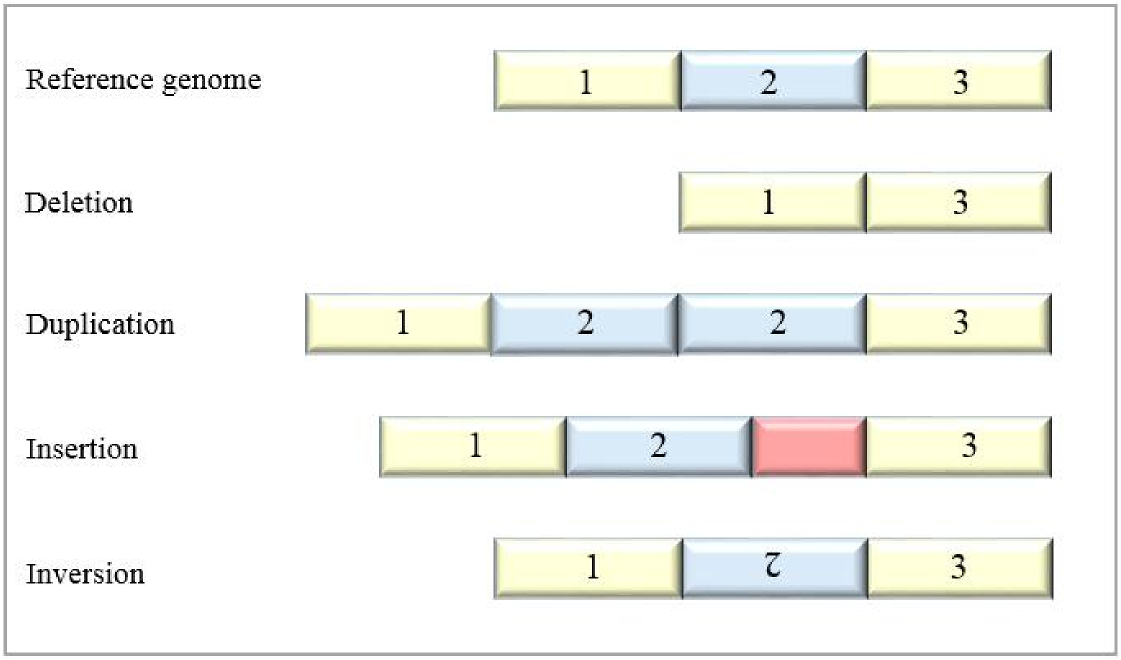
Tandem repeats or microsatellites are sequences of two to six nucleotides that are repeated several times along the DNA sequence. This figure illustrates the microsatellite repeat CTTCC sequence, repeated for two, three, or four times in different individual genomes.

More than one million microsatellite loci have been identified across the human genome [50,51]. The variable number of microsatellite repeats among individuals made them suitable genetic markers used in genetic studies, such as linkage analyses. The biological roles of microsatellites in disturbing genes, and thus influencing disease susceptibility, have been established for multiple diseases, such as Huntington disease and Fragile X syndrome [52,53].

1.2.3 Structural variations (SVs)

Due to the advances in DNA analysis technologies, recently other genetic variations in the human genome, such as the structural variations (SVs), were discovered [49]. SVs are DNA fragments that present with variable copy numbers among individuals (**Figure 1.3)** and include INDELs and CNVs. In addition to the genomes of patients with disorders, SVs have also been found in the genomes of healthy individuals [28,54]. Similar to SNPs, it has been estimated that SVs may play a significant role in population diversity,

Figure 1.3: Examples of structural variations (SVs).



complex behavioral traits, and disease characteristics that have not been explained by whole-genome SNP association studies [55].

SVs are distributed throughout the human genome. Mills et al (2006) noted that SVs may represent up to 25% of the genetic variations within the human genome, with an average of one variant for every 7.2 kbps [55]. More than 40% of the identified SVs were reported to be in coding and regulatory regions of the genes [31,49,56]. SVs are presumed to affect gene expression and function, and thus influence human phenotypes [56,57].

Multiple studies noted that SVs range in size from one bp to a few mega base pairs (Mbps) [55,58]. Feuk et al (2006) classified these variations into two groups: INDELs that are shorter than one kbp, and CNVs that range in size from one kbp to several Mbps

[49]. Throughout this thesis project, we annotate structural variations as INDELs or CNVs based on their sizes as described by Feuk et al (2006).

1.2.3.1 Insertion/deletion variants (INDELs)

INDELs are small (1-1,000 bps) insertions or deletions of DNA sequences in the genome [31,49]. INDELs have been estimated to be the second most abundant type of genetic variations (around 600,000), after SNPs [59,60].

INDELs can consist of multiple of three bps (3_n) or other numbers of bases [55]. The majority of the INDELs occur in the noncoding sequences; however, around 42% of them affect functionally important elements along the genome, and thus they may alter gene expression or function [59,61]. If 3_n INDELs arise within exon sequences, they result in the insertion or deletion of amino acids while maintaining the open reading frame [59]. On the other hand, a coding region INDEL consisting of a number of bases different than 3_n may lead to frameshift that would introduce a premature stop codon, resulting in a possibly non-functional gene [60,62]. INDELs that occur in the non-coding regulatory regions of genes have also been identified to affect gene expression and function. Therefore, not surprisingly INDELs have been identified to contribute to disease susceptibility and outcome risk, including in cancer [55,63,64]. For example, a study performed by Anderson et al (2010) on Danish colorectal cancer patients identified an INDEL in the nuclear factor of kappa B (*NFKB*) gene; this INDEL has been significantly associated with the elevated risk of colorectal cancer [65].

1.2.3.2 Copy number variants (CNVs)

CNVs are large deleted or duplicated DNA fragments that range in size from one kbp to several Mbps and exist in variable copy numbers among individuals [8,28,54,66,67]. If a CNV is found in > 1% of the individuals in a population, it is called copy number polymorphism (CNP) [68-70].

Most of the CNVs identified prior to 2004 were deletions or duplications that were implicated in high penetrant diseases, and consisted of Mbps of DNA sequences that were large enough to be visualized under the light microscope [28,54,71]. These CNVs overlapped with genes and affected phenotypes [8,72]. For example, a deletion of two or more of the alpha-globin genes has been reported to cause alpha-thalassemia [73,74]; a three Mbps deletion at 22q11.21 has been identified to cause the velo-cardio-facial syndrome [75]; and the duplication of chromosome 21 has been found to cause Down syndrome [76]. Because of their direct roles in disease development, these CNVs were classified as mutations.

CNV identification experiments were later extended to individuals without disease. The initial genome-wide CNV surveys of the human genome were performed in 2004 by Iafrate et al (2004) [54] and Sebat et al (2004) [28]. Despite the low resolution technology, incomplete ascertainment, and small numbers of individuals examined in these studies, hundreds of CNVs larger than 100 kbps were identified in the genomes of non-diseased individuals [28,54]. In 2006, Redon et al (2006) constructed the first whole genome CNV map and reported that CNVs cover around 12% of the human genome. A set of these CNVs was observed to affect genes functioning in biological pathways, such as cell adhesion, chemical stimulus, and neurohormone/neurotransmitter-related pathways [8].

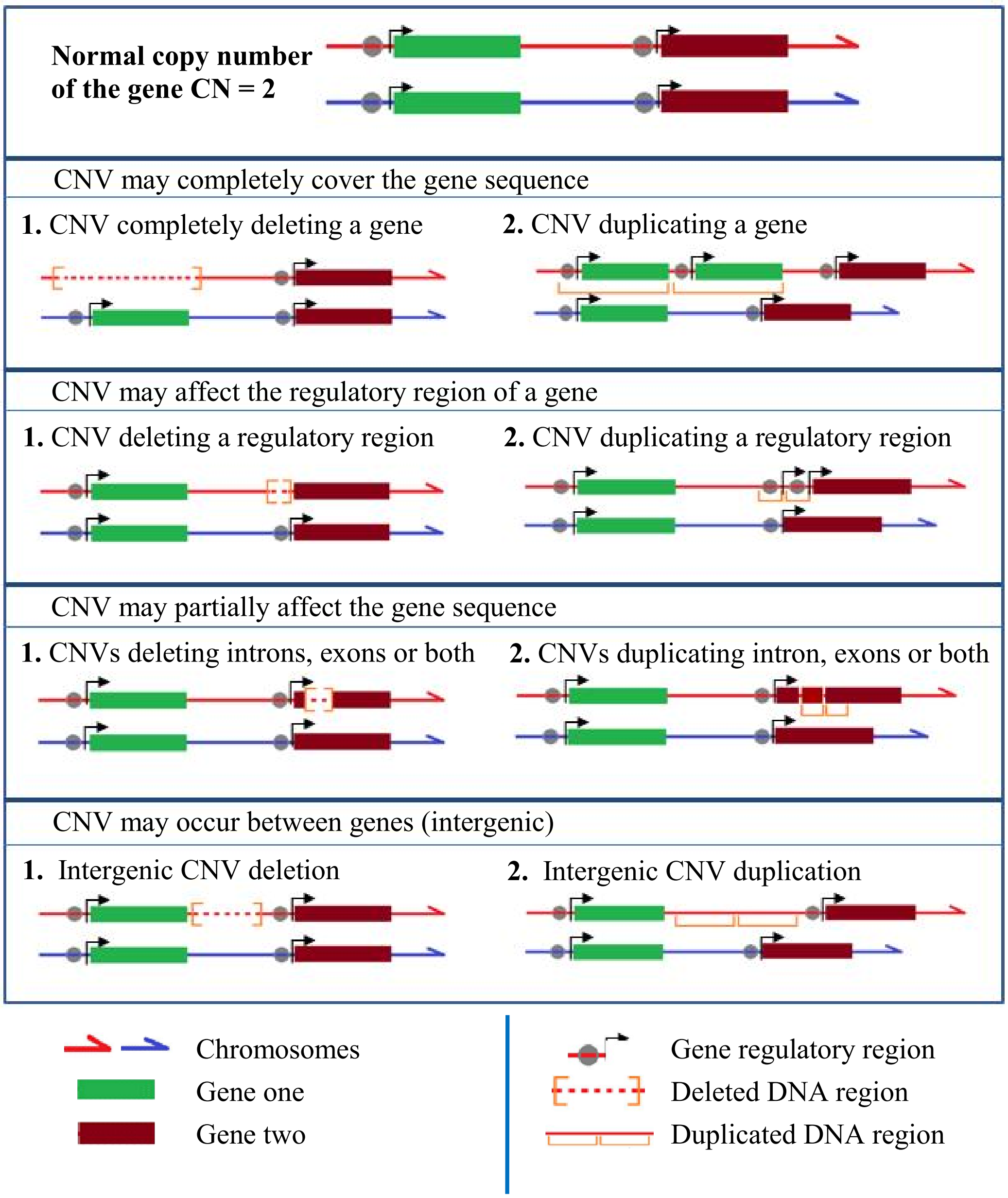
As the resolution of CNV identification technologies improved, the number of CNVs detected in the human genome has increased [77]. The majority of the identified CNVs have been observed to cluster in regions containing overlapping CNVs; these regions are called copy number variation regions (CNVRs) [8,78-80]. Recently, more than 15,000 CNVRs have been reported in the Database of Genomic Variants (DGV) [77,81-83]. These CNVRs are estimated to cover 10-15% of the human genome [77].

Multiple studies reported that nearly 99% of CNVs in the human genome are inherited, whereas the rest are caused by *de novo* mutations. Some *de novo* CNVs have clinical importance, since they have been identified to cause rare genetic disorders [84]. Apart from high-penetrant genetic disorders, some CNVs are also found to be associated with complex phenotypes and diseases as low-penetrant variations [7,8,66].

In previous studies, around 56% of the CNVs have been identified to overlap with genes [56,85]. Such CNVs may lead to different gene copy numbers and influence gene dosage, or cause abnormal gene sequence and structure, and thus intuitively modify gene expression or function [66,86].

CNVs can affect genes in different ways as shown in **Figure 1.4**. CNVs may completely delete or duplicate a gene or a group of genes. Deletion of genes lower the gene copy number and dosage, leading to down regulation of the affected genes. In contrast, duplication CNVs may increase the gene copy number and dosage, potentially causing over expression of the affected genes. However, not all CNVs that increase the gene copy number lead to an increase in gene dosage; Felekkis et al (2011) reported that the increase in gene dosages resulted by CNVs may be controlled by the action of miRNAs to keep the gene expression at the normal level [87]. Additionally, some CNVs

Figure 1.4: Possible ways by which CNVs can affect gene sequences.



This figure was adapted by permission from Cambridge University Press; Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med*. 2010;12:e8 [88]. **Appendix A** contains the copyright permission to use the figure.

partially cover genes [89]. If a CNV covers a gene's functional or regulatory region, it may cause up or down regulation of gene expression. Such CNVs may also lead to differential allelic expression of the affected genes [49,90,91]. It has been reported that the biological consequences of more than 50% of the identified CNVs in the genome is due to impact of CNVs on the regulatory regions of genes [92]. In some other cases, CNVs are located within a gene, and result in deletion or duplication of intron or exon sequences. CNVs affecting gene introns may influence the splicing sites, while CNVs affecting gene exons result in abnormal gene products [89]. Finally, some CNVs may occur within intergenic regions. Depending on their location, these CNVs may affect the regulatory regions of genes and may alter the gene expression and function [93]. These scenarios clearly depict that CNVs can contribute to population diversity and susceptibility to Mendelian and complex diseases [8,54].

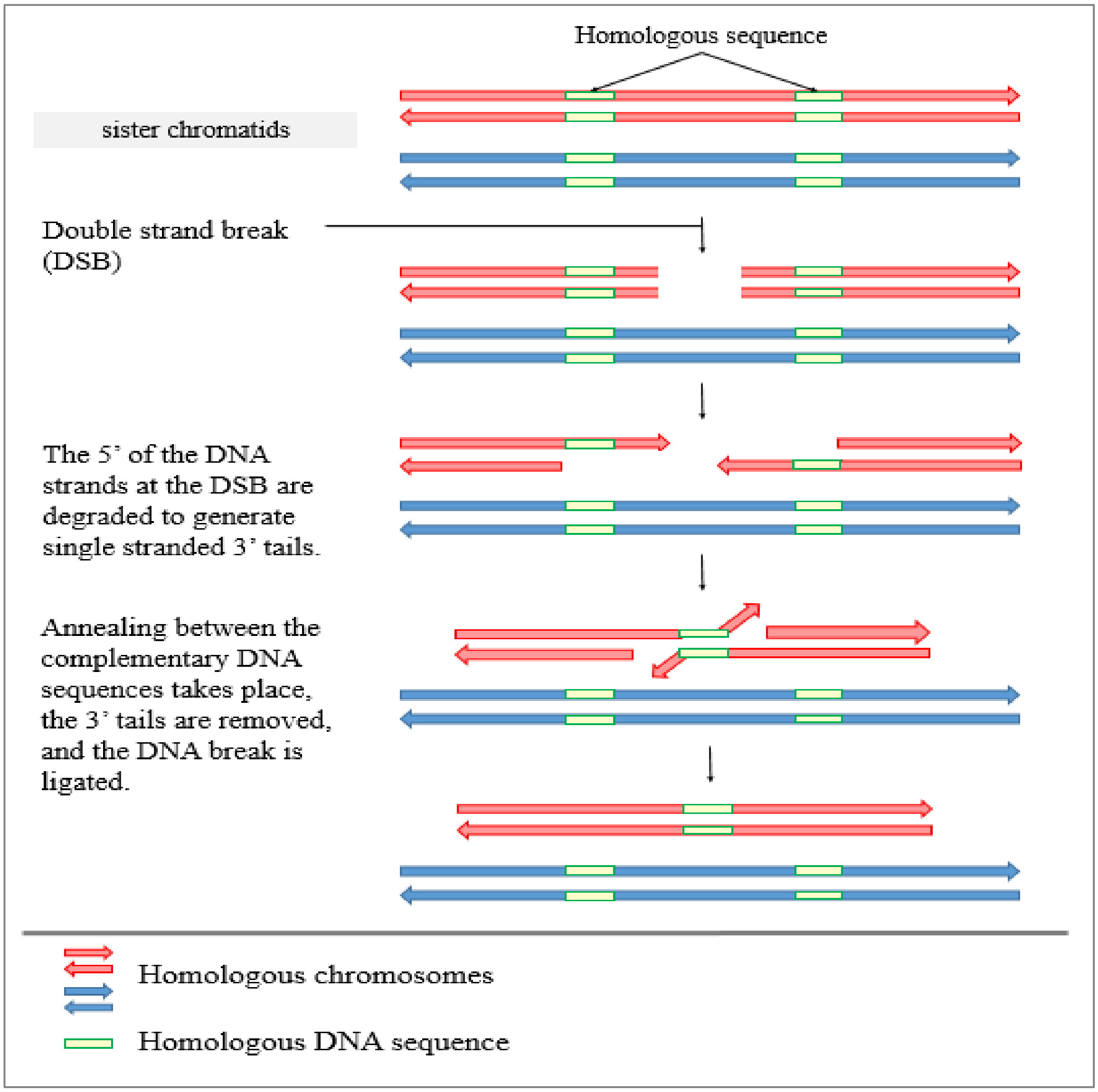
1.3 CNV Formation

Although the mechanisms of CNV formation are not completely understood at the present, several studies have estimated that DNA repair mechanisms play an essential role in the formation of structural variations [94,95]. During cell division and differentiation, various accidental lesions may arise in DNA. For instance, Double-Strand Break (DSB) is one of the fatal forms of DNA damage that may happen naturally during the cell cycle [96,97]. While DSB is required for meiotic recombination, it may also lead to genomic instability and cell death if it is not repaired by the DNA repair mechanisms [98]. Homologous recombination (HR) and non-homologous recombination (NHR) are the two major DNA repair pathways implicated in the formation of structural variations through DSB [95].

HR requires 300 bps of homologous DNA sequence as well as the activity of the Rad51 strand exchange protein to start repair of the DSBs [95,99]. Single strand annealing (SSA) is one of the HR mechanisms that repairs the DSB and may lead to deletion at the DSB [100] (**Figure 1.5**). In contrast to HR, NHR requires little or no homology in order to initiate the DNA repair [101,102]. NHR consists of non-replicative and replicative DNA repair mechanisms [95]. These mechanisms result in deletion, insertion, or inversion of DNA sequences that range from four bps to large regions, such as INDELs and CNVs [98,103-109]. For example, the nonhomologous end joining (NHEJ) mechanism rejoins the DSB without template sequence and leads to maximum of four bps deletion or insertion in the DSB site (**Figure 1.6**).

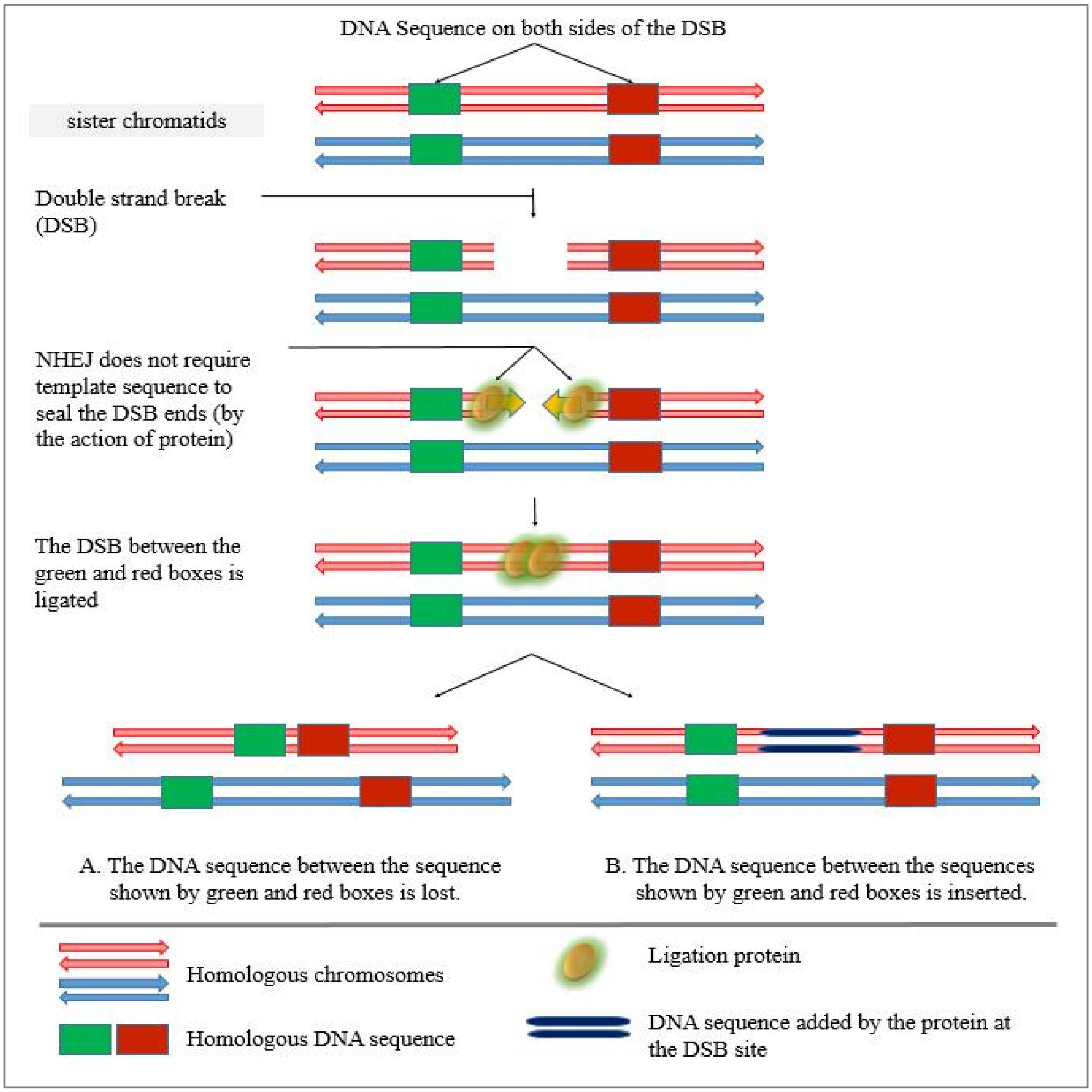
Multiple studies reported that segmental duplications (SegDups) are significantly associated with high rates of HR and chromosomal instability [110]. Not surprisingly, many literature reports linked SegDups to the formation of deletion, duplication, or inversion of DNA sequences, including CNVs [110-112].

Figure 1.5: Single strand annealing (SSA) mechanism.



SSA is one of the HR mechanisms that leads to the formation of CNVs. When both ends of the DSB cannot invade a nearby homologous repeat, the 5' of the DNA strands are degraded to generate single stranded 3' tails. As soon as a homologous sequence of at least 29 bps in both single stands is identified, annealing takes place at the complementary DNA by the action of the Rad52 protein, the 3' tails are removed, and the DNA at the break point is ligated. As a result, the DNA sequence located between the homologous DNA repeats (yellow boxes) is lost [95]. This figure is adapted by permission from Macmillan Publishers Ltd. [Nat Rev Gent]; Hastings P, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet*. 2009;10(8):551-564 [95]. The copyright permission is shown in **Appendix B**.

Figure 1.6: Nonhomologous end joining (NHEJ) mechanism.



NHEJ is one of the non-replicative mechanisms of NHR that does not require homology to repair DSB. This mechanism may rejoin DNA by: **(A)** leading to 1-4 bps deletions at the DSB site between the sequences shown by the green and red boxes, or **(B)** leading to insertion of random sequence shown by the dark blue region at the DSB site. This figure is adapted by permission from Macmillan Publishers Ltd. [Nat Rev Gent]; Hastings P, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet*. 2009;10(8):551-564 [95]. The copyright permission is shown in **Appendix B**.

1.4 Detection of structural variations

Several molecular and computational approaches have been used to detect SVs. These approaches include the fluorescent in situ hybridization (FISH) technique, bacterial artificial chromosome (BAC) arrays, array-based comparative genome hybridization (aCGH) technique, CNV arrays, and the whole genome single nucleotide polymorphism (SNP) arrays [54].

FISH is a cytogenetic technique that is traditionally used for detection of large chromosomal abnormalities. In FISH, fluorescently-labeled and single-stranded DNA probes complementary to specific genomic regions are spread on a glass slide. Then, the DNA sample of interest is denatured and applied on to the glass slide. Following the hybridization, the fluorescent signals are visualized to examine deletions or amplifications. While FISH has limited resolution (~5-10 Mbp) [113], it has been successfully used in the clinic to identify deletions, duplications, or translocations observed in many genetic disorders [114].

BAC array is a type of comparative genome hybridization (CGH) array. This technique identifies gross deletions or amplifications within the test genome in comparison to the reference genome [115,116]. BAC arrays were the first technique to be used for the genome-wide CNV analyses [8,54,117] . BAC array CGH uses segments of DNA from the sample under investigation along with a reference DNA that are inserted into bacterial plasmids. Then, the engineered plasmids are inserted into bacteria, such as *E. coli*, for replication [118,119]. Afterward, BAC molecules are isolated from bacteria, differentially labeled and hybridized with probes on solid surface arrays, such as glass slides. Then, the test and reference DNA signal intensities are recorded for all probes on the array. Significant deviation from the test/reference ratio of a probe or a series of consecutive probes would be interpreted as a DNA copy number change. BAC arrays have been used to investigate common chromosomal abnormalities in complex diseases, such as cancer [120,121], and diagnosis of many developmental disorders at an average resolution of about one Mbp [122].

Array-based comparative genomic hybridization (aCGH) is an upgrade of the traditional CGH technique. In aCGH, DNA samples can be obtained from different sources, such as BAC, polymerase chain reaction (PCR) products, or oligonucleotides. This array is used to compare CNVs of an individual under investigation with the genome of a reference individual [123]. DNAs from the test and the reference individuals are differentially labeled with fluorescent tags. Both DNA samples are then hybridized with the probes on a platform. Following this step, the differences in the fluorescent signals are used to identify the copy number gains (duplications) or losses (deletions) at each probe location [124]. Based on its resolution (10~25 kbps between the probes), aCGH produces data that is considered low to medium resolution [125]. Another limitation of aCGH is its limited ability to detect translocations and inversions along the human genome [126].

The high resolution probe-based arrays, such as CNV microarrays and SNP arrays are considered to be the most effective molecular technique for CNV detection; Pinto et al (2011) noted that the quality and quantity of detected CNVs increase significantly as the resolution of the detection array improves [77].

The high resolution human genome CNV microarrays are aCGH arrays developed to identify CNVs at the whole genome level or at targeted, specific genomic regions.

Genome-wide CNV arrays contain up to one million probes evenly distributed across the genome with average spacing of ~3 kb. Although, the CNV microarrays show a similar performance to aCGH microarrays, they are able to identify and discover small aberrations in segmental duplications and known CNVs [127] using empirically validated and optimized probes listed in the DGV [83].

SNP arrays were originally designed to genotype SNPs across the human genome, but the relatively high resolution of SNP arrays makes them also suitable for CNV detection and characterization [128]. During the SNP array reaction, single stranded DNA molecules are hybridized to hundreds of thousands of unique, fluorescently labeled probes. These probes are allele specific oligonucleotides that are complementary to the target regions along the DNA. After the primer extension reaction, the fluorescence intensity at each probe is detected by special scanners to make the SNP genotype calls and by further analyses to identify the CNVs [129].

Illumina® and Affymetrix are the two companies that offer probe-based SNP array platforms. Recent platforms produced by these companies include more than one million genetic markers to cover the human genome [128], including a group of CNV probes. These probes were carefully selected to cover CNVs that have been previously identified and experimentally validated [130]. Addition of the CNV probes to SNP genotype arrays improves the CNV detection using the SNP-array approach.

In addition to their ability of SNP genotyping and CNV identification, SNP arrays have multiple advantages over other classes of arrays. First, SNP arrays require less quantity of DNA per experiment than other platforms, such as aCGH. Second, in comparison to aCGH, SNP arrays are capable of detecting copy-neutral loss of heterozygosity (LOH) [130], which consists of duplication of one of the parental alleles and loss of the other at a specific locus [131]. Finally, the cost of SNP arrays is reasonable when compared to other techniques [130]. However, it is also worth mentioning that SNP arrays have some disadvantages when they are used for CNV detection. First, SNP arrays produce high noise-to-signal ratio, which requires data normalization [132]. Second, ~70% of the common CNPs have been found in SegDup regions and SegDup regions suffer from low SNP coverage. Therefore, identification of CNVs located in or around SegDups by SNP arrays is challenging [77,130,133]. This limitation of SNP arrays was considered during the design of the CNV microarrays, which detect CNVs throughout the whole genome including the SegDup regions [127]. Finally, while SNP arrays identify deletions more than duplications, CNV microarrays are optimized to detect deletions and duplication at the same rate [127,134].

1.4.1 Illumina® SNP BeadChip arrays and the hybridization technology

The latest generation of high density Illumina Infinium® HD platforms consist of dense (n ≈ 1.2 millions) and minimally spaced markers. These features increase the genomic coverage of the Infinium HD platforms, and facilitate a wide range of whole-genome DNA analysis, such as genome-wide SNP genotyping and structural variation detection analyses [135,136].

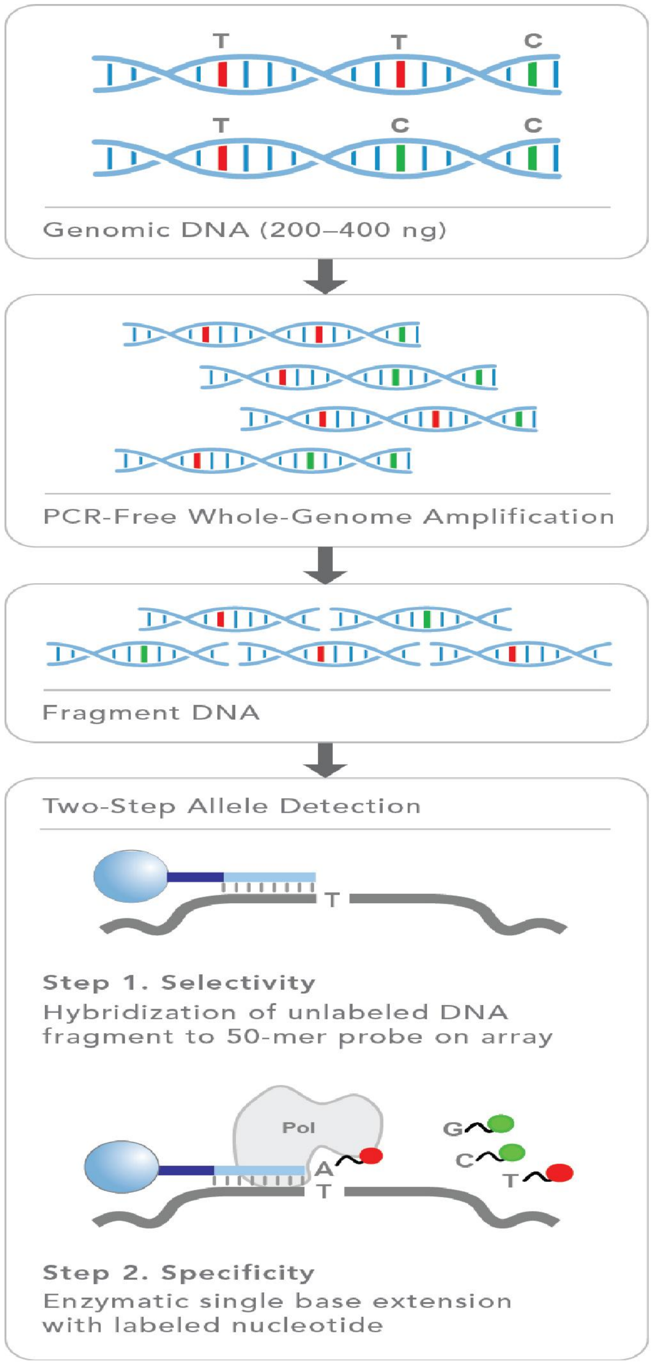
The four sample Illumina® Human Omni1 Quad platform is an example of the Illumina Infinium® HD platforms. This platform is regarded as one of the most powerful SNP genotyping arrays used in the GWAS analysis and CNV detections because of its features. For example, this platform is characterized by the high density of markers, with a median spacing of 1.2 kbps [137] and contains more than one million (n=1,134,514) selected probes. The majority of the probes (n=1,010,518) have been derived from the data produced by the HapMap Project [135,138]. Additionally, there are a total of 123,996 probes covering more than 11,000 common and rare CNVRs [77,139]. These CNV probes were selected in collaboration with the Centre for Applied Genomics at the Sick Kids Hospital in Toronto; the Wellcome Trust Sanger Institute, UK; and Harvard Medical School/Brigham and Women's Hospital, USA [77,139] based on new coding variants identified by the 1000 genomes project and other recent studies [140]. Additionally, the genome coverage has been estimated to be 93% for the Caucasian population (CEU), 92% for both the Han Chinese (CHB) and Japanese (JPT) populations, and 76% for the Yoruba in Ibadan (YRI) population [137]. Finally, the design of the platform enables the parallel genotyping of four DNA samples, decreases the amount of DNA required to 200 ng/sample, and provides the fastest and the most cost efficient platform for identification of disease causing/associated genetic variations [77,135].

Similar to other Illumina Infinium® HD platforms, the Illumina® Human Omni1 Quad technology comprises of 3µm silica beads that are self-assembled in 5.7 µm separated microwells on silica slides [141]. Each bead attaches to hundreds of thousands of probes; these probes are complementary to the sequences of specific genomic regions. Each probe consists of 80 nucleotides, in which the first 30 nucleotides at the 5’ end anneal to the bead, and the last 50 nucleotides are complementary to the DNA sequence adjacent to the targeted marker site [128].

With the high density Illumina Infinium® platform protocol, genotyping of a DNA sample is accomplished in four steps (**Figure 1.7**). First, the genomic DNA is fragmented. Second, DNA samples are applied on the array to hybridize with the probes. Third, labeled single nucleotides lacking the 3' hydroxyl group (dideoxynucleotide triphosphates; ddNTPs) are added to the reaction for a primer extension reaction [141]. When a ddNTP is added to the probe, it stops the DNA synthesis. The ddCTPs and ddGTPs are labeled with the biotin hapten stained with Alexa555 that results in green fluorescence, whereas the ddATP and ddTTP are labeled with the dinitrophenol hapten stained with Alexa647 that emits red fluorescence [129].

After these steps, the array is scanned by the iScan system for detection of the red and green fluorescence intensities at each marker [137]. This system reads the fluorescence intensity of each marker from two fluorescent intensity channels (X, Y), one channel per allele at the locus. The two fluorescent intensity channels provide intensity values called X_{raw} and Y_{raw} [137]. These raw intensity values are then imported into the Illumina Beadstudio® software for normalization [130] that is required to reduce the noise to signal ratio.

Figure 1.7: Illumina SNP array genotyping workflow.



This figure is courtesy of Illumina, Inc. The image was downloaded from the Illumina Image Store (http://www.illumina.com/company/news-center/multimedia-images.html) and used under the Illumina terms of agreement. Copyright agreement to use this image is shown in **Appendix C**.

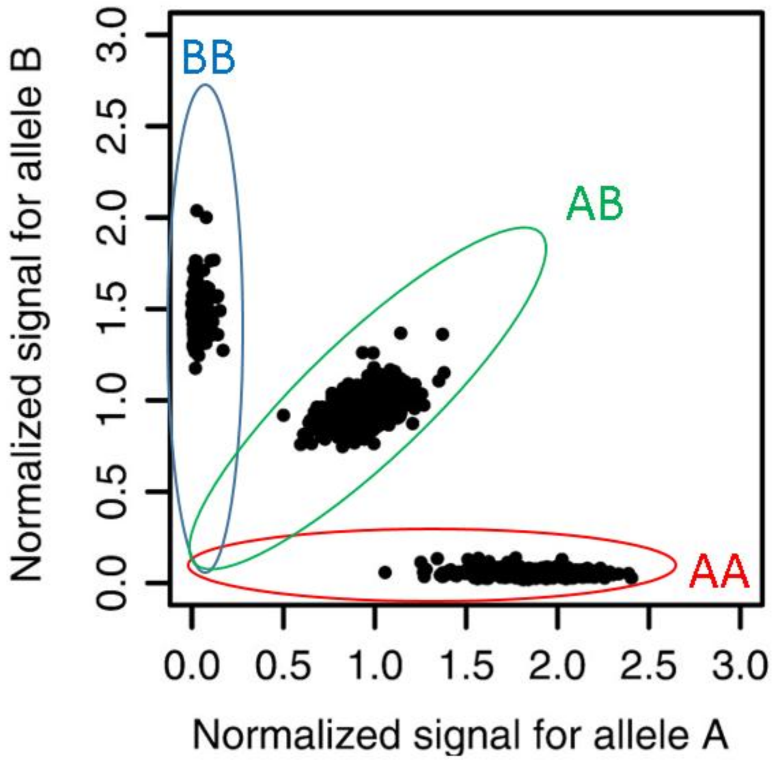
1.4.2 Raw intensity data normalization

Winchester et al (2009) [130] reported that SNP arrays have been optimized for SNP genotyping, and the noisy signals of the raw intensity data at the X_{raw} and Y_{raw} channels do not affect the accuracy of SNP genotyping; however, when these platforms are used for CNV detection, the noisy signals have been found to complicate the CNV calls [130,142]. Therefore, normalization of the array data is essential to adjust for a channel-dependent background and global signal intensity deviations. Usually, these deviations or variations in signal intensities are resulted from several factors, including human errors, differences in sample preparations, and variability in reagents [142-144].

As described elsewhere [145], the Illumina five step standard normalization algorithm, which is implemented in the Illumina Beadstudio® software, converts the raw signal intensities at X_{raw} and Y_{raw} to normalized intensity values. These normalized X_{norm} and Y_{norm} values denote normalized signal intensity at the A and B alleles of each marker.

Based on the three possible genotype categories (AA, AB or BB) at each SNP marker, Illumina Beadstudio® software identifies three cluster locations for each marker among the genotyped samples (**Figure 1.8**). These cluster locations are then used to identify the genotype calls, LOH, or Copy Number (CN) states [145].

Figure 1.8: Examples of genotype cluster positions.



The Cartesian plot of a SNP's genotype clusters: the X-axis shows the normalized intensity of A allele (red), and the Y-axis shows the normalized intensity of B allele (blue), whereas the intensity of the heterozygous genotype AB (Green) comes in the middle. This figure is modified from the Teo,Y.Y; Inouye,M.; Small,K.S.; Gwilliam,R.; Deloukas,P.; Kwiatkowski,D.P.; Clark,T.G.. A genotype calling algorithm for the illumina BeadArray platform. *Bioinformatics*. 2007;23(20):2741-2746 [146]. **Appendix D** contains the copyright permission to use this figure.

The normalized data is used to identify the total normalized intensity value of A and B alleles (R) and the allelic intensity ratio (Theta; θ) at each marker as follows:

- $R = X_{norm} + Y_{norm}$
- $\theta = \frac{2}{\pi} \arctan\left(\frac{Y_{norm}}{X_{norm}}\right)$

To identify CNVs using signal intensity data, two metric values are calculated from R and θ measures at each locus: Log_2 R ratio (LRR) and B allele frequency (BAF).

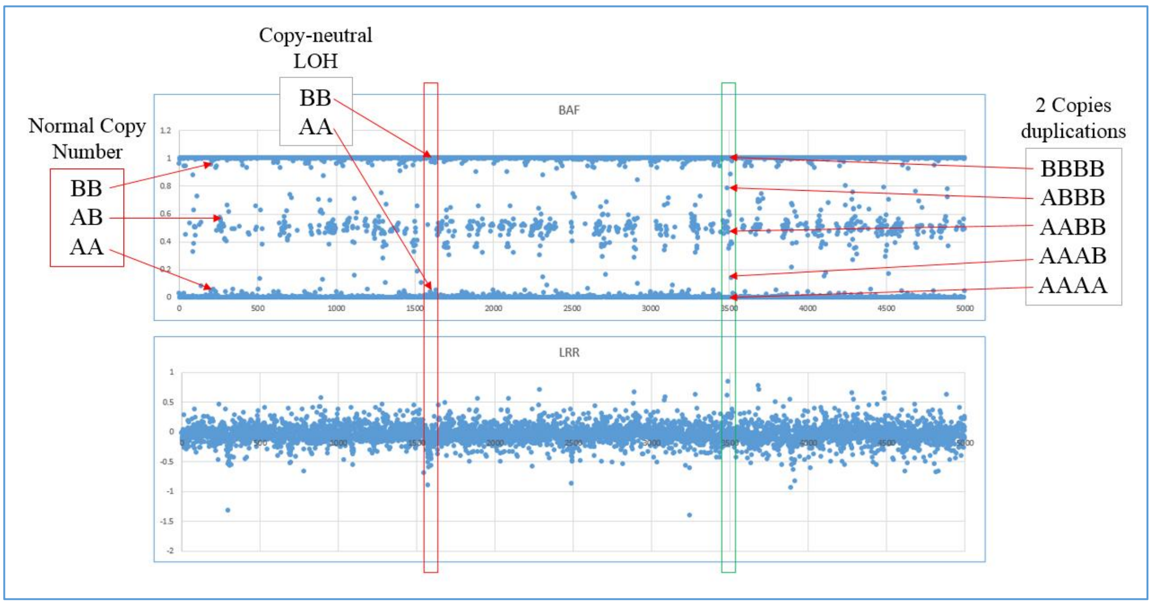
The LRR consists of the normalized total signal intensity at A and B alleles and is calculated as $\log_2 (R_{observed} / R_{expected})$, since the $R_{expected}$ is the R computed based on nearby genotype reference clusters. For example, if two alleles (A and B) are observed at a marker locus, then the $R_{observed} = 2$, and the $\log_2 2/2 = 0$ for the normal copy number of A and B alleles. On the other hand, if one allele (A or B) is deleted, only one allele is observed at the marker locus, thus the $R_{observed} = 1$, and the LRR is calculated for allele A or B as $\log_2 1/2 = -1$. In contrast, if the number of the (A or B) allele increases, for example an extra copy of the A or B allele is observed at the marker locus, then the $R_{observed} = 3$, and the LRR is calculated as $\log_2 3/2 = 0.58$. Based on the deviation of the LRR value from zero because of one copy deletion or duplication in this approach, it is obvious that deletions are predicted easier than duplications [134]. In fact, this finding can be considered as a limitation of using SNP arrays in CNV detection [77,147].

The second signal intensity measure is BAF, which is resulted from the allelic intensity composition that measures the percentage of A and B alleles at each locus. It also shows the theta value that identifies the deviation of the signal intensity of a locus from the cluster location. The BAF value is calculated by the following equation [18]:

- $$\text{BAF} = \begin{cases} 0, \text{if } 0 < \theta < \theta\text{AA}. \\ 0.5\frac{\theta - \theta\text{AA}}{\theta\text{AB} - \theta\text{AA}}, \text{if } \theta\text{AA} < \theta < \theta\text{AB}. \\ 0.5 + 0.5\ \frac{\theta - \theta\text{AB}}{\theta\text{BB} - \theta\text{AB}}, \text{if } \theta\text{AB} < \theta < \theta\text{BB}. \\ 1, \text{if } \theta\text{AA} < \theta < 1. \end{cases}$$

θ_{AA}, θ_{AB}, θ_{BB} represent the values of θ at the AA, AB and BB genotype clusters resulted from the study samples, respectively. For example, for a sample having a normal copy number, the frequency of the B allele at AA, AB, BB genotypes equal to 0.0, 0.5, and 1.0, respectively. **Figure 1.9** illustrates the LRR and BAF signal intensity values for different CN states.

Figure 1.9: Illustration of BAF and LRR values at a normal, deleted, or duplicated CN state.



Plot of the BAF and LRR signal intensity values for a selected region of chromosome 6 (created based on the study data). The normal chromosome region contains three BAF genotype clusters (AA, AB, and BB) genotypes, and with LRR values arranged around zero. In the copy-neutral LOH region, BAF has (AA and BB) genotypes, but it lack the AB genotype, however this region has normal LRR values. The increased copy number region shows increased LRR values and an increased number of peaks in the BAF distribution. LOH: loss of heterozygosity.

1.4.3 CNV detection using the Illumina® SNP genotyping data

Several studies reported that the choice of a CNV calling algorithm can be as important in the accuracy of CNV detection as the choice of the array used. It is well documented in the literature that results from different algorithms may differ in terms of the quality and quantity of the CNVs called [77,78,130].

Traditional approaches of CNV identification comprise of the examination of signal intensities; these approaches compute the mode of BAF for SNPs in a sliding window approach along the chromosome to detect copy number changes. These models are simple to use, and yet the sliding window approaches have a limited ability to identify the exact CNV borders [18]. Therefore, detection of CNVs from high resolution platforms and accurate identification of CNV breakpoints have required the development of robust techniques and sophisticated calling algorithms [132].

Most of the recently developed algorithms are based on either hidden Markov models (HMM) or segmentation approaches. Sample-based CNV calling algorithms, also called non-segmenting algorithms, process each sample independently based on the appropriate clustering file (e.g., HapMap samples) with canonical cluster positions for each SNP; for example, QuantiSNP [17] and PennCNV [18] are the most popular, published sample-based CNV calling algorithms used to detect CNVs using the Illumina® SNP genotyping platform data. Both QuantiSNP and PennCNV are based on a HMM in which the LRR and the BAF are considered independently as observed states at any given marker locus, and they are used in the HMM to detect the hidden state of the number of copies [148,149]. Both of these algorithms were developed and optimized to detect CNVs from Illumina SNP genotyping arrays, and they were preferred to be used in CNV research because of their prediction sensitivity and breakpoint identification accuracy [77,150-154]. These two algorithms were used in this thesis project, and are described in detail in the following sections.

1.4.3.1 QuantiSNP

QuantiSNP is a computational CNV calling algorithm, which uses an Objective Bayes hidden-Markov Model (OB-HMM) to infer and detect CNVs from BeadChip SNP genotyping data [17,77]. OB-HMM has been found to significantly improve the resolution of CNV detection, in which OB measures are used to set hyperparameters that calibrate the model to a fixed false positive rate. Additionally, the HMM uses LRR and BAF values to infer the status of the unknown copy number (hidden state) at each locus based on the most preceding marker [17]. The changes of one copy number state between neighboring SNPs describe the probability of shifting from one state to another [130]. There are six hidden states detectable by the QuantiSNP algorithm (**Table 1.1)**.

Table 1.1: The copy number states used by the QuantiSNP algorithm.

Hidden state	Copy number (CN)	CNV Genotypes
1	0	Null
2	1	A or B
3	2	AA, AB, or BB
4	2 (LOH)	AA or BB
5	3	AAA, AAB, ABB, or BBB
6	4,5	AAAA, AAAB, AABB, ABBB, or BBBB

CNV hidden states, copy number states, and interpretation of the CNV geneotypes used by the QuantiSNP algorithm. This table is modified from Colella et al, QuantiSNP: An objective bayes hidden-markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35(6):2013-2025 [17]. This table was used under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/). LOH: loss of heterozygosity.

QuantiSNP also compares the evidence of the presence of a copy number variance in a region versus the normal state (two copies). This step is performed based on the data using an Expectation Maximization (EM) algorithm. As a result, QuantiSNP computes a

Bayes Factor (BF) at each detected copy number variation region and reports the log BF as a confidence value. As the BF value increases, the probability of a CNV existence increases [17,153].

1.4.3.2 PennCNV

PennCNV is another academically designed, published, and freely available CNV detection algorithm, which uses an integrated HMM model to predict CNVs using the Illumina® BeadChip SNP genotyping array data [18]. PennCNV uses a combination of values including LRR and BAF values, the population frequency of the B allele (PFB), and the distance between neighboring markers that helps determine the probability of copy number state transition between the adjacent markers. These values are all used by the first-order HMM [77,153] during the CNV prediction.

The first-order HMM of PennCNV infers the hidden copy number state at any given locus based on the copy number state of the most adjacent marker, unless a transition in copy number states is detected between neighboring markers [18,130]. This HMM also uses LRR and BAF values for each locus to develop models for transition between different copy number states, as well as to differentiate between the neutral copy number LOH regions and the normal state regions [18]. Furthermore, PennCNV uses the family trio information (when available) to improve the CNV prediction and accuracy of boundary mapping, as well as to detect novel CNVs [18].

Compared to many other CNV detection algorithms that use loss, normal and gain terms to demonstrate the copy number state, PennCNV uses six copy number states of

CNVs (**Table 1.2**). In contrast to QuantiSNP, PennCNV assigns the total copy number 4 for variations with duplication of two copies and above [18].

Table 1.2: The copy number states used by the PennCNV algorithm.

Copy number state	Total copy number	CNV genotypes
1	0	Null
2	1	A or B
3	2	AA, AB, or BB
4	2 (LOH)	AA or BB
5	3	AAA, AAB, ABB, or BBB
6	4	AAAA, AAAB, AABB, ABBB, or BBBB

CNV hidden states, copy number states, and the interpretation of CNV genotypes used by the PennCNV algorithm. This table is modified from Wang et al., PennCNV: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007;17(11):1665-1674 [18]. This table was used under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC-BY-NC). LOH: loss of heterozygosity.

PennCNV was originally designed to detect CNVs from Ilumina® BeadChip data, but it has been recently modified to predict CNVs from other platforms as well, such as the Affymetrix platforms [77].

It should also be mentioned that while both QuantiSNP and PennCNV algorithms are developed to detect CNVs, they are also capable of detecting INDELs.

1.4.3.3 The advantages of using more than one CNV detection algorithm

As briefly mentioned above, CNVs are inferred using the SNP array data based on the signal intensity measures resulted from the hybridization of fluorescence probes. Carter et al (2007) reported that hybridization studies usually generate false positive and false negative predictions, which should also be assessed while using SNP genotyping arrays to detect CNVs [132]. To overcome this problem in CNV detection, most of the recent CNV detection studies suggested using more than one algorithm in CNV prediction to increase the accuracy of the results [130,155-157] and the breakpoint identification, and to decrease the false positive prediction rate [77,78,132,153]. Specifically, if a CNV is detected by more than one algorithm, the confidence in the existence of this CNV increases; although such an approach also may increase the false negative predictions and lead to elimination of true variations. Accordingly, the latest CNV studies have used at least two CNV calling algorithms to predict CNVs.

While this is a well-accepted idea, it also creates a challenge: it is not unusual for a CNV to be detected with un-identical boundaries by different algorithms. That is why an "overlap" analysis is required where CNVs detected by different algorithms in the same region are tested for a certain amount of shared sequence, and if the length of the shared sequence exceeds a pre-defined threshold, then these CNVs are assumed to be the same CNV detected by multiple algorithms [78,155,156]. In this thesis project, I used the threshold of at least 50% overlap [155,158] to consider a CNV to be predicted by both QuantiSNP and PennCNV (**Section 2.6.3.2**).

1.5 CNVs in human diseases

Due to their relatively new discovery and importance, studies on CNVs in relation to human health and disease are rapidly emerging. Studies published so far have demonstrated the biological and medical importance of CNVs in susceptibility and outcome in many diseases, such as autism, schizophrenia, diabetes, asthma, and cancer [86,159].

CNVs have been also reported to contribute to the risk, progression, or survival of cancer [90]. Notably, deletions may down regulate the expression of tumor suppressor genes, and amplification can up regulate the expression of oncogenes. For example, Jin et al (2011) identified a common 32 kbps deletion CNV (CNP2454) on chromosome 20p13 that affects the Signal-Regulatory Protein Beta 1 (*SIRPB1*) gene and significantly associates with aggressive prostate cancer [160]. Additionally, a deletion of more than 272 kbps at 3p22.3, which affects 5 of the 33 exons of the deleted in lung and esophageal cancer 1 (*DLEC1*) gene was reported to possibly predispose individuals to lung and esophageal cancers [161].

CNVs may also have protective roles; for example, a duplicated copy of a CNV affecting the human CC chemokine ligand 3-like 1 (*CCL3L1*) gene has been reported to be significantly associated with low susceptibility to the human immunodeficiency virus (HIV) infection and acquired immunodeficiency syndrome (AIDS) [162].

While the role of CNVs in human phenotypic variability and disease characteristics becomes increasingly evident, a complete understanding requires additional studies in many other human conditions. For example, in the case of colorectal cancer, which is studied in this thesis project, the role of CNVs in altered gene function or expression and the disease susceptibility or prognosis remains largely unknown.

1.6 Colorectal cancer

Colorectal cancer is the neoplasm of cells in the innermost mucosa layer of the colon or the rectum. More than 90% of colorectal cancer cases are sporadic adenomas [163], whereas the rest of disease cases are due to inherited mutant genes [164]. Colorectal cancer does not have detectable symptoms at its early stages, and as a result it is usually diagnosed at advanced stages [165]. The symptoms of the disease include occult or asymptotic anemia, bright red or dark blood in the stool, abdominal discomfort and change in the bowel movements, anorexia, weight loss, nausea, vomiting, and tiredness [165-167].

1.6.1 Incidence and mortality rates of colorectal cancer

Colorectal cancer is a major health problem worldwide: it is estimated to be the third most frequent type of cancer and the fourth leading cause of cancer-related deaths [1]. According to the assessments of a study published in 2008, colorectal cancer is estimated to account for 9.4% of the total cancer cases worldwide (nearly 1.2 million cases and 608,000 deaths) [168,169].

The incidence and mortality rates of this disease vary among nations; the incidence rates are identified to be higher in developed countries when compared to developing countries, presumably due to diet and a westernized lifestyle [1,170,171]; the highest colorectal cancer rates have been reported in Europe, Japan, Australia, New Zealand and North America [169]. Conversely, the lowest rates have been estimated in South Central Asian, Central and South American countries, and some parts of Africa [1].

In Canada, colorectal cancer is considered as one of the most significant health problems. Based on the Canadian Cancer Statistics 2014 report generated by the Canadian Cancer Society, about 24,400 new colorectal cancer cases were expected to be diagnosed during 2014 [3]. Colorectal cancer is estimated to be the second most common cancer in males, with 1:13 males predicted to be diagnosed with colorectal cancer during their life time [3]. In females, colorectal cancer is estimated to be the third most frequent type of cancer. The life time risk of Canadian females to develop the disease is 1:16 [3].

While the mortality rates of colorectal cancer have been declining steadily since 2003 as a result of screening and treatment improvement in Canada, colorectal cancer remains to be the second and third cause of cancer related deaths for males and females, respectively [3]. Unfortunately, the highest incidence and mortality rates of colorectal cancer among the Canadian provinces have been identified in Newfoundland and Labrador (NL) [3].

1.6.2 Pathology of colorectal cancer

The majority of colorectal cancer cases arise from polyps in the innermost mucosa layer of the large intestine [172]. Polyps are benign, abnormal growth of cells that may develop within the large intestine. Polyps can occur in different sizes and shapes; however, the large polyps elevate the risk of cancer formation [173]. There are two types of large intestine polyps: hyperplastic polyps and adenomatous polyps. Hyperplastic polyps that arise from epithelial tissue are not pre-cancerous polyps, and most of them remain benign and rarely transform to cancer, even though doctors prefer to remove them if they are detected during screening [174]. However, adenomatous polyps (adenomas) that arise from the glandular tissue of the mucous membrane are called pre-cancerous polyps, as they may transform into cancer [164,173]. These polyps represent about two-thirds of the large intestine polyps. While less than 10% of them transform to cancer, more than 95% of colorectal cancers arise from adenomatous polyps [175].

1.6.3 Risk factors of colorectal cancer

Although not all causes of colorectal cancer are known at the present, there are specific factors that have been identified to modify the risk of this disease. These factors consist of a combination of environmental and genetic factors.

Diet is one of the important factors contributing to the risk of colorectal cancer. So far, consumption of red meat and animal fat, low intake of fruits, vegetables, and fiber have been associated with the increased risk of colorectal cancer [176,177]. Cigarette smoking and regular alcohol consumption have also been estimated to elevate the risk of colorectal cancer [178].

Obesity is another factor that may modify the risk for the colorectal neoplasia. Multiple studies have identified associations between obesity and the development of adenomas. For instance, individuals having a body mass index (BMI) greater than 30 were estimated to have 1.5-2.4 fold higher risk of colorectal cancer when compared to people having BMI less than 22 [178].

Multiple studies noted that lack of physical activity is also correlated with the increased risk of colorectal cancer. A meta-analysis done by the World Cancer Research Fund estimated that regular exercise and physical activity lead to a statistically significant risk reduction (15%) of adenoma formation [179]. Another study suggested that 12-14% of colon cancer cases can be due to low level of physical activity [180].

Other known colorectal cancer risk factors include age, gender, ethnicity, and genetic factors. It has been identified that more than 90% of colorectal cancer cases occur after the age of 50, and the susceptibility to this disease elevates significantly after this age [166]. Gender is another risk factor for colorectal cancer, since men show a slightly higher risk to develop colorectal cancer than women [181].

Based on the worldwide distribution of the incidence of colorectal cancer, it has been suggested that ethnicity may also have a role in colorectal cancer development. Black populations are found to have a higher susceptibility to colorectal cancer in comparison to white populations, while Asian and Pacific Island populations have the lowest risks of colorectal cancer when compared to other ethnic groups [182]. However, Asians may have increased risk to develop colorectal cancer, if they migrated to high risk countries [183] , due to the western diet.

Family history is another important colorectal cancer risk factor. About 25% of the colorectal cancer cases are reported to have a positive family history of colorectal cancer, and individuals having one or more relatives diagnosed with the disease in different generations are at increased risk to develop colorectal cancer [184,185]. Furthermore, the risk of developing this disease has been found to be two fold higher if a first degree relative, such as a parent, sibling or offspring, has been diagnosed with colorectal cancer. Therefore, similar lifestyle or shared genetic compositions between the members of families are the likely features contributing to familial colorectal cancer [186].

1.6.4 Genetics of colorectal cancer

Lynch et al (2003) reported that high penetrant genetic mutations account for 5-10% of all colorectal cancer cases worldwide [187]. These mutations have been reported to underlie the familial and inherited colorectal cancer syndromes, such as familial adenomatous polyposis (FAP), hereditary non-polyposis colon cancer (HNPCC), Juvenile polyposis (JPS), Peutz-Jeghers syndrome (PJS) and others [185]. On the other hand, around 90% of colorectal cancer cases are sporadic that are assumed to be resulted from a combination of low penetrant genetic variations and environmental factors [188].

1.6.4.1 Familial and inherited colorectal cancer

A. Familial adenomatous polyposis (FAP)

Familial adenomatous polyposis (FAP) is an inherited autosomal dominant form of colorectal cancer that accounts for 1% of the cases [164,189]. FAP is caused by inherited germline mutations in the adenomatous polyposis coli (*APC*) gene. *APC* is a tumor suppressor gene located on chromosome 5q21 and mutations of the *APC* gene account for 95% of all colorectal adenomatous polyposis cases [190].

FAP patients initially develop hundreds to thousands of benign adenomatous polyps; the large number of these polyps increases the probability of later developing invasive tumors [191]. Individuals diagnosed with FAP have a 100 % risk of developing colorectal cancer with an estimated median diagnosis age of 40 years [189].

A milder type of FAP is called attenuated familial adenomatous polyposis (aFAP). AFAP is associated with the mutations in the 5’ end (in exon 9), and in the distal 3’ regions of the *APC* gene. AFAP is characterized by the fewer number of polyps and delayed onset of colorectal cancer (~12 years) when compared to FAP [190].

Turcot syndrome is a rare type of FAP that is also caused by mutations in the *APC* gene; however, it may also be caused by mutations in the mismatch repair genes, mutL homolog 1 (*MLH1*) and postmeiotic segregation increased 2 (*PMS2*) genes associated with Lynch syndrome [192]. Individuals diagnosed with Turcot syndrome develop many polyps in the colon or rectum, in addition to the tumors in the brain and spinal cord [190,193].

B. Lynch syndrome

The second well described type of hereditary colon cancer is Lynch syndrome, which was often referred to hereditary nonpolyposis colorectal cancer (HNPCC). About 1-3% of all colorectal cancer cases are due to Lynch syndrome. This disease is an autosomal dominant condition that occurs in families and may lead to multiple members being diagnosed with colorectal or other types of cancer. The mean age of the diagnosis of Lynch syndrome is estimated to be 44 years [164].

Lynch syndrome is caused by the germline mutations in one of the four DNA mismatch repair (MMR) genes; mutS Homolog 2 (*MSH2*), mutS Homolog 6 (*MSH6*), *MLH1*, and *PMS2* genes [190,194]. Mutations in these MMR genes lead to microsatellite instability in the genome [164]. It has been noted that individuals carrying mutations of MMR genes have ~35% lifetime risk to develop colorectal cancer [195]. Also, they may have an increased risk of developing a wide range of extra-colonic malignancies [190], such as endometrial cancer, adenocarcinomas of the stomach, malignant tumors of the small bowel, hepatobiliary tract, ovary, upper urinary tract, and pancreas, as well as glioblastoma of the brain [190,196].

Muir-Torre syndrome (MTS) is considered to be a subtype of Lynch syndrome that are associated with glioblastomas and sebaceous skin tumors [197]. MTS have been associated with mutations in MutL homolog 3, mismatch repair (*MLH3*), Postmeiotic Segregation Increased 1, mismatch repair (*PMS1*), and Transforming Growth Factor, Beta Receptor II (*TGFBR2*) genes [198,199].

Colorectal cancer cases that show the characteristics of HNPCC, but do not have mutations in the MMR genes, are classified as another type of HNPCC called familial colorectal cancer type X (FCCX). Although the genetic etiology of FCCX remains unknown, multiple studies estimated a possible association of the variations in the bone morphogenetic protein receptor type 1A (*BMPR1A*) gene [200], or the mutations in the ribosomal protein S20 (*RPS20*) gene [201] with the risk of developing FCCX.

C. Juvenile polyposis syndrome (JPS)

Juvenile polyposis syndrome (JPS) is an autosomal dominant condition caused by the mutations in the mothers against decapentaplegic homolog 4 (*SMAD4*) and *BMPR1A* genes [202,203]. JPS is characterized by the development of hamartomatous benign polyps in the colon or rectum, which have a potential to transform into cancerous tumors [164,189]. Additionally, individuals having JPS develop other forms of cancer, such as pancreatic, lung, breast, uterine, ovarian, and testicular cancers [202].

D. Peutz-Jeghers syndrome (PJS)

Peutz-Jeghers syndrome (PJS) is another autosomal dominant syndrome that is caused by mutations in the Serine/Threonine Kinase 11 (*STK11*) tumor suppressor gene, which regulates cell polarity and proliferation [204,205]. PJS patients develop hamartomatous polyps and cancerous tumors in the gastrointestinal tract, in addition to the formation of dark blue or brown freckles inside the mouth, face, fingers, or toes [189,206].

E. Other rare inherited colorectal cancer disorders

Rare inherited colorectal cancers include MUTYH-associated polyposis [207,208], hyperplastic polyposis syndrome (HPPS) and Cowden's syndrome (CS) [209,210].

1.6.4.2 Sporadic colorectal cancer

Sporadic colorectal cancer represents more than 90% of all colorectal cancer cases worldwide [185,188]. This type of colorectal cancer occurs in individuals who do not have a family history of the disease. In sporadic colorectal cancer, the effects of a group of low penetrant alleles combine with the effects of several environmental factors, such as diet and physical activity, to form hyperplasia and then adenoma [211].

GWASs identified a group of colorectal cancer susceptibility loci. These genetic markers include the germlin1 (*GREM1*); bone morphogenetic protein 2 (*BMP2*); mother against decapentaplegic homolog 7 (*SMAD7*); colorectal cancer associated 2 (*LOC120376*); colorectal cancer associated 1 (*FLJ45803*); chromosome 11 open reading frame 53 (*c11orf53*); POU class 2 associating factor 1 (*POU2AF1*); POU class 5 homeobox 1 pseudogene 1 (*POU5F1P1*); V-myc avian myelocytomatosis viral oncogene homolog (*MYC*); small subunit processome component homolog (*c8orf53*); eukaryotic translation initiation factor 3 subunit H (*EIF3H*); bone morphogenetic (*BMP4*); cadherin 1 (*CDH1*); and rho GTPase binding protein 2 (*RHPN2*) genes [189,212-215].

1.6.5 CNVs and colorectal cancer

Multiple studies reported chromosomal changes in the non-tumor genomes of colorectal cancer patients [216-218]. For instance, some CNVs, such as insertions at 7p14.1, 7p15.3, 7q31.2, 8q22.11, 3q31.3, and 13q32.4 and deletions at 17p13.1, 18q21.33, and 15q26.1 have been estimated to influence the expression of tumor suppressor genes (TSGs) and oncogenes in the microsatellite stable hereditary non-polyposis colorectal cancer (MSS HNPCC) [219].

Germline CNVs have been also estimated to alter the expression and function of a group of genes that code for proteins acting in multiple biological pathways, such as the Wnt pathway (transcription factor 7-Like 2, *TCF7L2*), chromatin-remodelling (tet methylcytosine dioxygenase 2,3; *TET2* and *TET3*), receptor tyrosine kinases (including receptor tyrosine-protein kinase erbB-3; *ERBB3*) , cell cycle checkpoint kinase (ataxia telangiectasia mutated, *ATM*), multiple fusion transcripts (including insulin-like growth factor 2; *IGF2*), fusions involving R-spondin family members (R-spondin 2,3; *RSPO2* and *RSPO3*), and a tumor suppressor gene (*PPMIL*). These genes have been identified to play important roles in colorectal cancer development and prognosis [220,221]. Interestingly, some CNVs were also reported to improve the survival of patients. For example, a 1.1 kbp long deletion has been found to delete the entire exon 4 of the mitochondrial tumor suppressor gene 1 (*MTUS1*) located on chromosome 8p21.3-22. This deletion CNV has been identified to increase the gene activity and decrease the disease progression in a wide range of human cancers [222-224], including colorectal cancer [225].

1.7 Rationale and research objectives

CNVs are biologically important genetic variations that are shown to contribute to population diversity and development of many diseases. However, as newly discovered variants, their exact roles or relations to many traits and diseases, such as colorectal cancer, are yet to be fully understood.

The main objectives of this thesis project were:

1) to computationally predict and characterize the germline genome-wide INDEL/CNV profiles in a cohort of colorectal cancer patient (n=505);
2) to identify the genes and biological pathways that may be impacted by the predicted INDELs/CNVs.

To my knowledge, this research project is one of the first genome-wide studies in colorectal cancer. The unique and comprehensive data generated during this study therefore significantly contribute to our understanding of the nature and biological relevance of INDELs/CNVs to colorectal cancer. Additionally, the generated data will be used in the Savas lab for other studies, such as examining the relation of INDELs/CNVs to colorectal cancer progression and survival outcomes. By being utilized in the future studies in the Savas lab or other labs around the world, this data will therefore be indispensable for further scientific discoveries related to INDELs/CNVs and colorectal cancer.

Chapter 2: Materials and Methods

2.1 Ethics approval

This research study was approved by the Health Research Ethics Authority (HREA) of Newfoundland and Labrador (HREA Reference #: 13.073).

2.2 Contributions and credits

Salem Werdyani: (Memorial University of Newfoundland) performed the computational analyses of this study. In particular, handled, processed and analyzed large-scale genomic data using a variety of bioinformatics tools and programming languages such as Java and Perl; downloaded the required programs, created the necessary data files and utilized QuantiSNP [17] and PennCNV [18] algorithms to detect the INDEL and CNV profiles in the genomes of the study cohort; used a set of quality control (QC) parameters and inclusion/exclusion filtering to reduce the false positive findings; used multiple statistical tools and computational programs, such as PLINK [226] to execute summary statistics and to describe the predicted INDELs/CNVs in term of their size, frequency, and CN state; identified the genes that are possibly affected by the predicted variations and performed biological pathway analysis to predict the potential biological consequences of INDELs and CNVs; interpreted the study results.

Dr. Wei Xu: (Princess Margaret Hospital, University Health Network, Toronto, Ontario) led the QC, inclusion/exclusion filtering, and population structure analyses for the study cohort after the SNP genotyping reactions.

Jingxiong Xu and **Konstantin Shestopaloff:** (Princess Margaret Hospital, University Health Network and University of Toronto, Ontario, respectively) contributed to the QC, inclusion/exclusion filtering, and population structure analyses for the study cohort under the supervision of Dr. Wei Xu.

Dr. Roger Green and **Dr. Patrick Parfrey:** (Memorial University of Newfoundland and NFCCR) provided the genome-wide signal intensity data for the patients included in this thesis project.

Dr. Jane Green and **Dr. Elizabeth Dicks:** (Memorial University of Newfoundland and NFCCR) contributed to patient recruitment and data collection.

Dr. Sevtap Savas: (Memorial University of Newfoundland) designed, supervised, and led the study; helped interpret the results; provided the baseline table for the study cohort based on the NFCCR data.

Funding agencies

This study was mainly funded by Colon Cancer Canada.

2.3 Patient cohort

This thesis project is conducted in cooperation with the NFCCR. A total of 750 colorectal cancer patients from Newfoundland were recruited to the NFCCR between Jan 1, 1999 to Dec 31, 2003. These patients were younger than 75 years at the time of diagnosis [227]. Patients were asked to provide blood samples and permission to access their tumor tissues and medical reports, and filled in questionnaires after they or their close relatives were consented [15]. A total of 539 patients with available genomic DNA as well as clinicopathological and outcome data were included into a genome-wide SNP genotyping experiment, outputs of which were used in this project.

2.4 Genome-wide SNP genotyping reaction

DNA samples were genotyped using the Illumina® Human Omni1_Quad_v1 genome-wide SNP genotyping array by a service provider (Centrillion® Biosciences, CA, USA). As it was previously stated in **Section 1.4.1**, this platform is a high resolution Illumina Infinium® BeadChip that provides genome-wide SNP genotype and signal intensity data, including Log R ratio (LRR) and B Allele Frequency (BAF) values for 1,134,514 genetic markers, including 1,010,518 SNP probes and 123,996 CNV probes. The data that was produced by this platform reported the marker positions based on the human genome coordinate 19 (Hg19).

2.5 Initial QC and population structure analyses using the genotype data

After the genotyping experiments, Dr. Wei Xu and his team of researchers performed the primary QC, inclusion/exclusion filtering, and population structure analyses for the patient cohort based on the genotype data [16]. As a result, individuals were excluded from the patient cohort if they; a) had discordant sex information (n=1); b) were 1^{st}, 2^{nd} or 3^{rd} degree relatives of another patient in the cohort (n=21); c) had outlying heterozygosity rate (n=1); or d) were non-Caucasian (n=11). After these QC and exclusion analyses, 505 out of 539 patients constituted the study cohort [16]. The baseline characteristics of the 505 patients are summarized in **Table 2.1.**

Table 2.1: The baseline characteristics of the 505 colorectal cancer patients of this study.

Features	Number	%
Sex		
Female	198	39.21
Male	307	60.81
Age at diagnosis	median: 61.43 years (range: 20.7-75 years)	
Location		
Colon	334	66.14
Rectum	171	33.86
Histology		
Non-mucinous	448	88.71
Mucinous	57	11.29
Stage		
I	93	18.42
II	196	38.81
III	166	32.87
IV	50	9.90
Grade		
Well/moderately differentiated	464	91.88
Poorly differentiated	37	7.33
Unknown	4	0.79
Vascular invasion		
Absent	308	60.99
Present	159	31.49
Unknown	38	7.52
Lymphatic invasion		
Absent	298	59.01
Present	167	33.07
Unknown	40	7.92
Familial risk		
Low risk	250	49.50
Moderate/high risk	255	50.50
MSI status		
MSI-L/MSS	431	85.35

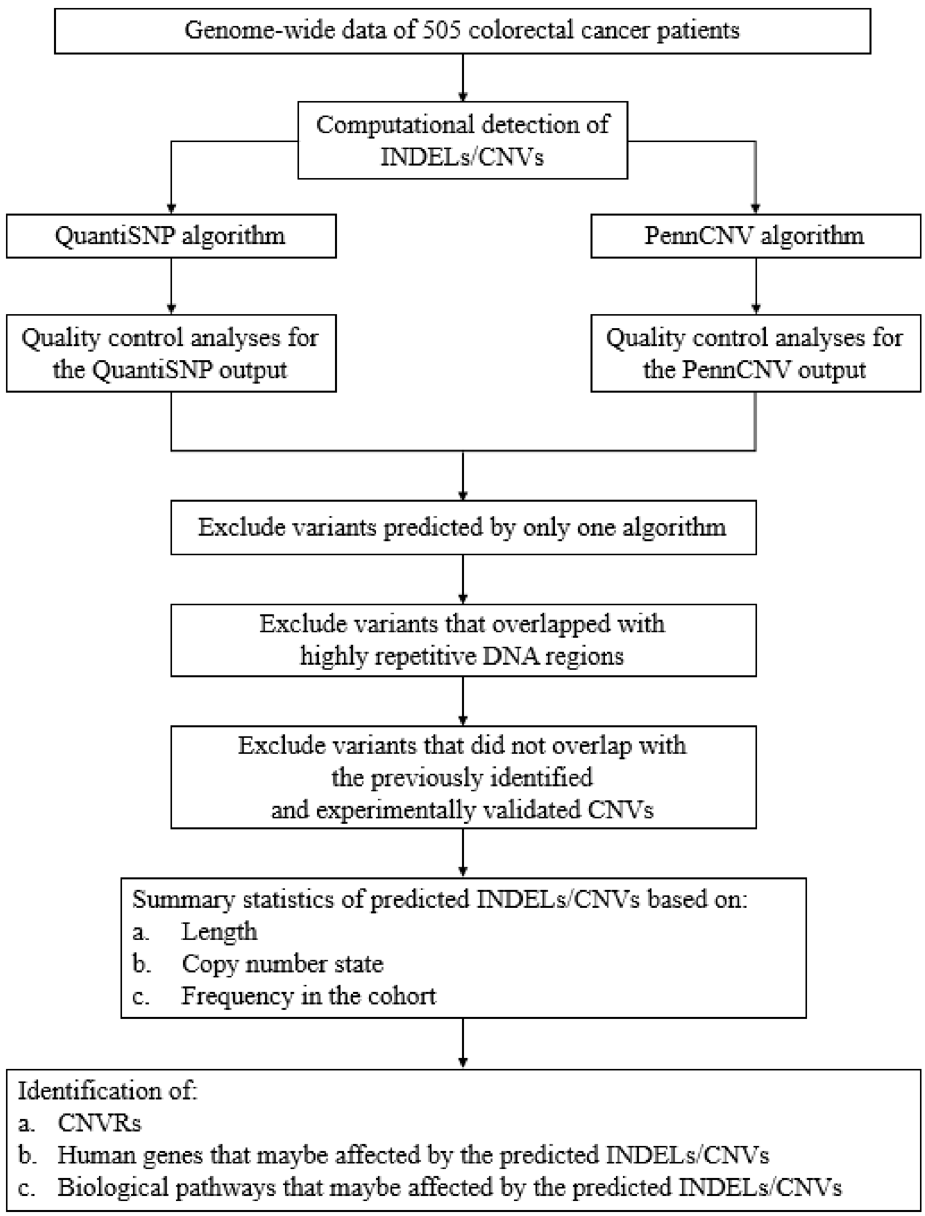
MSI-H	53	10.49
Unknown	21	4.16
Tumour *BRAF* Val600Glu mutation		
Absent	411	81.38
Present	47	9.31
Unknown	47	9.31
Colorectal cancer cases		
Sporadic cases	475	94.06
Lynch syndrome cases	14	2.77
FCCX cases	13	2.57
FAP cases	3	0.60

MSI-H: microsatellite instability-high; MSI-L: microsatellite instability-low, and MSS: microsatellite stable; FCCX: familial colorectal cancer type X.

2.6 Computational analyses of INDELs/CNVs

Prediction and description of INDEL/CNV profiles were performed in a series of stages as summarized in the flowchart shown in **Figure 2.1**.

Figure 2.1: The main stages of the study that were used to predict and describe the INDELs/CNVs in the patient cohort.



2.6.1 Computational detection of INDELs/CNVs

As discussed in **Section 1.4.3.3**, the majority of the recent CNV studies recommend using more than one CNV detection algorithm to increase the accuracy of the breakpoint estimation and to decrease the false positive finding rate [77,82,130,132,155-157]. That is why the computational prediction of INDELs/CNVs in the patient genomes was performed using two algorithms, QuantiSNP [17] and PennCNV [18]. These two algorithms are designed and optimized to detect CNVs from the whole genome SNP genotyping platform data.

The first step in the INDEL/CNV prediction was to generate the signal intensity files for each patient. Signal intensity files contain marker names, chromosome numbers, marker positions, and the signal intensity values (LRR and BAF) for 1,134,514 markers in the genotyping platform. For this study, the signal intensity files were created by merging two types of data files that were provided by the genotype service provider using a custom Perl program. These data files were; a) report data files, which include signal intensity data (LRR and BAF values) obtained during the genotyping reaction at each marker, and b) the final report MAP file that includes chromosome numbers, marker names, and marker positions based on the human genome assembly Hg19. These files, when merged together, become the signal intensity files suitable to be used by QuantiSNP [17] and PennCNV [18] algorithms.

Similar to other studies [228,229], due to the complexity of the analysis of the sex chromosome data, in this thesis project the prediction of INDELs/CNVs by QuantiSNP and PennCNV algorithms was performed for the autosomal chromosomes only.

2.6.1.1 Prediction of INDELs/CNVs by QuantiSNP

QuantiSNP algorithm [17] has been developed by the Wellcome Trust Centre for Human Genetics at Oxford University, UK. This algorithm had been optimized to detect CNVs from the genome-wide signal intensity data obtained during SNP genotyping reactions. To predict INDEL/CNV profiles, QuantiSNP (version 2) package was downloaded on April 19, 2013 from the QuantiSNP download website [230]. The signal intensity files of each patient were then used as input files to detect INDELs/CNVs by QuantiSNP using the default parameters.

Variable GC contents exist among genomic regions. These GC differences may lead to "genomic waviness" in the signal intensity values and complicate the computational detection of INDELs/CNVs [231]. Therefore, during this study, to correct for the fluctuation of the GC content in signal intensity measures, a GC correction step was also performed using QuantiSNP, as recommended by other studies [232].

2.6.1.2 Prediction of INDELs/CNVs by PennCNV

PennCNV [18] is the second CNV detection algorithm used in this project to predict INDELs/CNVs from the genome-wide signal intensity data of patients. PennCNV algorithm has been developed by Dr. Kai Wang and his colleagues in the Department of Genetics, University of Pennsylvania, USA [18]. For this thesis project, PennCNV algorithm and all of its supporting programs and additional required data files were downloaded on May 1, 2013 from the PennCNV download website [233].

In addition to the signal intensity files, other input data files are required to run the PennCNV algorithm. These files are the Population Frequency of B allele (PFB) and the GC-model file [233].

The PFB file contains the frequency of the B allele for each marker in the population that is required during the PennCNV analysis. PFB file also provides PennCNV with the chromosome coordinate information for each marker during the analysis [18]. While PFB file provided with the PennCNV algorithm package was generated based on the human genome assembly Hg18 (Genome Reference Consortium, GRCh36), the signal intensity data of the study cohort was generated based on the Hg19 genome coordinates (GRCh37). Thus, during this study a new PFB file based on the Hg19 genome coordinates was generated as follows; first, an Illumina® Human Omni1_QuadV1 dataset containing the signal intensity files for 88 HapMap CEU (Caucasian) individuals was downloaded on May 7, 2013 from the Gene Expression Omnibus database (GEO) database [234]. These signal intensity files were created by Illumina® based on the Hg18 genome coordinates and were uploaded to the GEO database under platform number (GPL8882) and series (GSE17197). Second, the (HumanOmni1-Quad_v1-0_B-H_MappingInformation.txt) file, which includes the Hg18 genome coordinate information and their equivalent for the Hg19 genome coordinates, was downloaded on January 14, 2014 from the Illumina® support website [235]. This mapping information file was used to substitute the Hg18 genome coordinate information with the Hg19 information in the 88 HapMap CEU signal intensity files using custom Perl programs. Finally, the reformatted 88 HapMap CEU signal intensity files were used to generate the PFB file using the Perl program *Compile_PFB.pl* that is provided within the PennCNV package [233].

Similar to the QuantiSNP analysis, while using PennCNV correction of genomic waviness in the signal intensity data was previously recommended [231,236,237]. To correct for the GC content, a GC-model file including the GC content of 500 kbps upstream and downstream of each marker was required [231,237]. Since the GC-model file provided with the PennCNV package was based on the Hg18 genome coordinates, during this study a GC-model file based on the Hg19 genome coordinates was also created. This procedure required two data files. The first is the *GC5Base.txt* file that contains the percentage of the GC bases in 5-base windows based on the Hg19 genome coordinates; this file was downloaded on January 28, 2014 from the University of California Santa Cruz (UCSC) genome bioinformatics download website [238,239] as suggested elsewhere [233]. Second, the signal intensity file of a randomly selected patient was used in the Perl program *Cal_gc_snp.pl* that was provided with the PennCNV package [233] to generate the GC model file based on the Hg19 genome coordinates.

After the generation of the required input files (signal intensity files, the PFB file and GCmodel file based on the Hg19 genome coordinates), INDELs/CNVs in the patient genomes were predicted by the PennCNV algorithm using the default parameters in the Perl program *Detect_cnv.pl* [233].

It has been noted by the PennCNV developer that if a high density SNP array is used to generate the signal intensity data, the PennCNV algorithm tends to split large CNVs into smaller ones [18]. Hence, similar to other studies [18,240,241], adjacent CNVs were merged together in the present study if the sequence gap between them did not exceed 1/2 of the total distance from the start position of the first CNV to the end position of the second CNV. This was done using the *Clean_cnv.pl* program of the PennCNV package once [233].

2.6.2 Post-prediction QC analyses

During the variant detection process, both QuantiSNP [17] and PennCNV [18] algorithms create QC files for each subject and each predicted INDEL/CNV. For example, QuantiSNP algorithm creates QC values, such as BAF standard deviation (BAF_SD), LRR standard deviation (LRR_SD) and a confidence score [17]. Similarly, PennCNV algorithm generates a set of QC values consisting of LRR_SD, B allele frequency drift (BAF_Drift), waviness factor (WF), absolute GC waviness factor (|GCWF|), BAF median, and a confidence score [18]. Appropriate threshold levels for these parameters were identified by an extensive literature search [77,78,82,130,153,154,156,157,237,242], which were then used in this study as explained below.

2.6.2.1 QC analyses for QuantiSNP outputs

To perform the QC analyses for INDELs/CNVs predicted by QuantiSNP, a custom Perl program was developed and the QC criteria were applied to both subjects and predicted INDELs/CNVs based on the selected QC parameters (**Table 2.2**).

Table 2.2: Exclusion criteria for the subjects and INDELs/CNVs based on the QuantiSNP QC data.

Exclusion Criteria		Threshold	References
Subject filtering	LRR Standard Deviation (LRR_SD)	> 0.28	[77]
	BAF Standard Deviation (BAF_SD)	> 0.20	[82]
	INDEL/CNV number per sample	> Mean + 3 SD	[240]
	Samples with extremely long CNVs	> 7.5 Mbps	[77,240]
INDEL/CNV filtering	INDEL/CNV length	< 10 bps	[78,157]
	Number of probes per INDEL/CNV	< 10 probes	[237,241]
	Confidence Score (Max Log Bayes Factor)	< 30	[78,242,243]

LRR: Log R Ratio, **BAF:** B Allele Frequency, **SD:** Standard Deviation.

2.6.2.2 QC analyses for PennCNV outputs

The PennCNV algorithm package [233] has a Perl program (*filter_cnv.pl*) that was used in this study to perform the QC analyses using the criteria outlined in **Table 2.3**.

Table 2.3: Exclusion criteria for the subjects and INDELs/CNVs based on the PennCNV QC data.

Exclusion Criteria		Threshold	References
Subject filtering	LRR Standard Deviation (LRR_SD)	> 0.28	[153]
	BAF drift	> 0.01	[153]
	LRR waviness factor (WF)	≤ - 0.04 and ≥ 0.04	[152,233]
	BAF median	< 0.45 or > 0.55	[158,244]
	INDEL/CNV number per sample	> Mean + 3 SD	[77,240]
	Samples with extremely long CNVs	> 7.5 Mbps	[77,240]
INDEL/CNV filtering	INDEL/CNV length	< 10 bps	[78,157]
	Number of probes per INDEL/CNV	< 10 probes	[237,241]
	Confidence score	<10	[82,153,158,245]

LRR: Log R Ratio, **BAF:** B Allele Frequency, **SD:** Standard Deviation.

2.6.2.3 Summary statistics of the INDELs/CNVs predicted by QuantiSNP and PennCNV algorithms

Following the QC filtering, INDELs/CNVs predicted by QuantiSNP and PennCNV algorithms were examined in more detail. In these analyses, PLINK statistical tool [226] was utilized to define the predicted INDELs/CNVs based on their lengths and CN states. Three PLINK input files called CNVlist, MAP, and FAM were utilized during this step.

The list of the INDELs/CNVs predicted by QuantiSNP was converted to the PLINK input format (the CNVlist file) using a custom Perl program, while the list of variations predicted by PennCNV was converted to the PLINK input format by using the Perl supporting program *PennCNV_to_ PLINK.pl* [233]. In addition, the MAP file used by PLINK during this step contained the standard genotype MAP information, where the dummy markers represented the start and end positions of each predicted INDEL/CNV. Finally, the FAM file included the family, gender, and phenotype data of each patient. After the preparation of the PLINK input files, they were loaded in PLINK to identify the number of variation per individual, variants' length, and their CN state.

2.6.3 Filtering the predicted INDELs/CNVs

During this study, in addition to the QC analyses, further inclusion/exclusion filtering steps were performed. These analyses aimed to identify INDELs/CNVs that were: a) predicted by one algorithm and found to overlap with each other in the same genome; b) predicted by both algorithms and identified to overlap with each other in the same genome; c) overlapped with the highly repetitive DNA regions; and d) overlapped with the experimentally identified CNVs.

2.6.3.1 Identification of INDELs/CNVs predicted by one algorithm and overlapped with each other in the same individual

If two INDELs/CNVs are found to overlap with each other in the same individual's genome, they are estimated to be one variant mistakenly detected twice by a CNV detection algorithm [232]. In order to check this possibility in my data, PLINK [226] was utilized to investigate the predictions made by QuantiSNP and PennCNV separately. To do so, similar to **Section 2.6.2.3**, the list of predicted INDELs/CNVs were converted into the PLINK input file format and based on this data, new MAP and FAM files were created. These three files were then loaded into PLINK, which assessed the overlap between INDELs/CNVs in the genome of each patient.

2.6.3.2 Identification of the overlapping INDELs/CNVs predicted by both the QuantiSNP and PennCNV algorithms in the same individual

As it has been mentioned in **Section 1.4.3.3**, if the signal intensity data was generated by SNP genotyping platforms, most of the CNV studies suggest using more than one CNV detection algorithm to identify CNVs [130,132,155,156]. The reason for that is if INDELs/CNVs are detected by two or more algorithms with the same CN state and have a certain portion of their lengths overlapping with each other, they are most likely to be the same variant [78,155,156,158,246,247]. In addition, prediction of a variant by more than one algorithm increases the confidence in accuracy of the predictions [77,130]. These points were considered in this study. Since one bp overlap criteria may be too relax (which would increase the false positive predictions), or 100% length overlap criteria may be too strict (which would increase the false negative findings), similar to a number of other studies [155,248] in this study I opted for a reasonable criterion of at least 50% length overlap; in other words, the variants that were predicted by both algorithms with the same CN state and at least 50% length overlap were assumed to be the same variant.

To do so, a Perl program was written to identify the possible overlaps between the variations predicted by both algorithms with the same CN state in the same genome. Possible ways of overlap between variations are shown in **Figure 2.2**. This program merged the variations together if they have the same CN state and had at least 50% of their length overlapped in the same genome. The boundaries, and therefore the length of the merged variations, were estimated based on the smallest (downstream) start position to the largest (upstream) end position of the overlapping variations. The length of the merged variant is also called "the union length" [77,157,246], which was used to identify the percentage of overlap between the overlapping variations using the Jaccard Similarity Coefficient [77]. This coefficient was calculated in this Perl program as follows:

$$\text{Overlap percentage (\%)} = \frac{\text{Intersection length} * 100}{\text{Union length}}$$

Then the Perl program used the length, percent overlap, and the CN state to determine and exclude the INDELs/CNVs predicted by only one algorithm, predicted with different CN states (for example, an INDEL/CNV was excluded if it was identified

Figure 2.2: Possible ways of overlap between INDELs/CNVs predicted by the QuantiSNP and PennCNV algorithms.

1	PennCNV (Start 1) X--------------------------------X (End 1) QuantiSNP (Start 2) X--------------------------------X (End 2) Identical start and end positions by both algorithms
2	PennCNV (Start 1) X-------------------------------X (End 1) QuantiSNP (Start 2) X----------------------------X (End 2)
3	PennCNV (Start 1) X-----------------------------X (End 1) QuantiSNP (Start 2) X----------------------------X (End 2)
4	PennCNV (Start 1) X--------------------------------X (End 1) QuantiSNP (Start 2) X-----------------X (End 2)
5	PennCNV (Start 1) X-------------------X (End 1) QuantiSNP (Start 2) X-----------------------------X (End 2)
6	PennCNV (Start 1) X-----------------------X (End 1) QuantiSNP (Start 2) X--------------------------------X (End 2)
7	PennCNV (Start 1) X--------------------------------X (End 1) QuantiSNP (Start 2) X-----------------------X (End 2)
8	Penn CNV (Start 1) X--------------------------------X (End 1) QuantiSNP (Start 2) X----------------------X (End 2)
9	Penn CNV (Start 1) X----------------------X (End 1) QuantiSNP (Start 2) X--------------------------------X (End 2)

as a deletion by one algorithm and a duplication by the other in the same patient), or had less than 50% of their union length overlapping with each other.

Following this analysis, PLINK [226] was utilized to describe the general features of the overlapping INDELs/CNVs in terms of their lengths and CN states. As mentioned in **Section 2.6.2.3**, PLINK analysis requires PLINK input files that included these new variants' data (CNVlist) and generation of the corresponding MAP and FAM files. These files were first generated and then were loaded in PLINK for analysis.

2.6.3.3 Exclusion of INDELs/CNVs overlapping with the highly repetitive DNA regions

Signal intensity data derived from the highly repetitive DNA sequences, such as centromere and telomere regions, leukocyte Immunoglobulin-like receptor gene cluster and olfactory receptor (OR) genes may complicate the CNV detection [18,245,249]. For this reason, INDELs/CNVs that overlapped at least one bp with these DNA regions were excluded from further analyses.

To perform this analysis, a list of highly repetitive DNA regions based on the Hg19 genome coordinates was generated through multiple steps as follows; a) the genome coordinate information for leukocyte Immunoglobulin-like receptor gene cluster based on (Hg18) and centromere positions based on (Hg19) was obtained on February 1, 2014 from the PennCNV website [233]; b) the LiftOver tool of the UCSC genome browser [239] was then used on February 4, 2014 to change the genome coordinates of the leukocyte Immunoglobulin-like receptor gene cluster from Hg18 to Hg19 genome coordinates; c) the UCSC genome browser [239] was utilized on February 11, 2014 to identify the start and end positions of each chromosome based on Hg19. Then the telomere regions were determined by adding and subtracting 500 kbps at the start and end positions of each chromosome respectively, as suggested in the PennCNV package [233]; d) the list of centromere positions provided by the PennCNV website (see "a" above) was adjusted by adding and subtracting 100 kbps to upstream and downstream of each centromere following the PennCNV recommendations [233]; e) a full list of OR genes based on the Hg19 genome coordinates was downloaded on February 14, 2014 from the Human Olfactory Receptors Data Explorer (HORDE) database [250]. As a result, information for 840 autosomal OR genes from the HORDE database was obtained. **Appendix E** contains the list of centromere and telomere regions, leukocyte Immunoglobulin-like receptor gene cluster and OR genes based on the Hg19 genome coordinates.

Finally, a new Perl program was formulated to identify and exclude the INDELs/CNVs having at least one bp overlap with these highly repetitive DNA regions.

2.6.3.4 Identification of INDELs/CNVs overlapping with the experimentally validated CNVs

Several studies estimated that INDELs/CNVs overlapping with the previously identified and experimentally validated variations are most likely to be existing (i.e. true) CNVs, but not methodological artifacts [78,156,158,251]. Therefore, the INDELs/CNVs that were predicted in this study and had at least 50% of their length overlapping with the previously identified CNVs were also identified.

On April 1, 2014, a list of previously identified CNVs based on the Hg19 genome coordinates was downloaded from the DGV database website [83]. This dataset contained three published studies [31,252,253] that identified the INDELs/CNVs in large numbers of DNA samples (n=451-1,414) using experimental methods, such as CGH, oligo CGH, or whole genome sequencing approaches. These studies reported 29,202 variants that were used to compare with the INDELs/CNVs predicted in this project with the help of Perl programming. The resulted INDELs/CNVs were deemed to be the final list of variations that were predicted with high-confidence (high-confidence variants).

Similar to the previous stages (**Section 2.6.2.3**), for this set of variants summary statistics analyses were performed by PLINK [226].

2.7 Identification of distinct, high-confidence INDELs/CNVs

Once the final list of high-confidence variants was identified, a Perl code was written to determine the INDELs/CNVs that exist in different individuals within the study cohort. This program checked for the INDELs/CNVs that had the exact start and end positions among the subjects, and created a list of distinct, high-confidence INDELs/CNVs that were detected at least once within the study cohort. The frequency of each distinct INDEL/CNV was also calculated during this step. INDELs/CNVs found in less than 5% of the patients were considered rare variants and INDELs/CNVs found in at least 5% of the patients were classified as common variants. Information related to the length of the INDELs/CNVs and the detected CN states in the patient genomes were also obtained.

2.8 Identification of CNVRs

It is known that some genomic regions (CNVRs) contain multiple CNVs [8,246]. Therefore, similar to other studies [8,78-80], the distinct high-confidence INDELs/CNVs identified in **Section 2.7** were examined to estimate the CNVRs in the data set. In order to perform this analysis, an additional Perl code was used to identify at least one bp overlap between the distinct, high confidence INDELs/CNVs predicted in this study.

2.9 Identification of the genes possibly affected by the INDELs/CNVs

Nearly 56% of the previously identified human CNVs have been reported to overlap with genes [85,91]. To identify the gene sequences that may be overlapping with the distinct INDELs/CNVs predicted in this study, the genomic coordinates for human expressed sequences (based on the Hg19) were downloaded on August 15, 2014 from the ENSEMBL database [254]. This list was then filtered to have the information for autosomal genes only, which was used to identify the overlap between the distinct INDELs/CNVs and the human genes. This analysis was performed by using Perl programing.

2.10 Identification of the biological pathways that may be affected by the INDELs/CNVs

Since proteins functionally interact with other proteins in biological networks, the biological pathway analysis was performed to interpret the data in the context of gene function, biological processes, pathways, and networks. PANTHER database [255], which recently integrated the information by the Gene Ontology (GO) annotations database [256], was utilized to identify the biological pathways that are possibly affected by the predicted INDELs/CNVs. Specifically, the list of the ENSEMBL genes identified in **Section 2.9** was loaded into the “Gene List Analysis” tool of the PANTHER database [255] on September 5, 2015, which then returned the pathway information for the genes.

Chapter 3: Results

In this study, genome-wide signal intensity data were used to predict the genome-wide INDEL and CNV profiles of 505 colorectal cancer patients. The computational prediction of the INDELs/CNVs was performed by utilizing two CNV detection algorithms. A set of QC and inclusion/exclusion filtering was then performed to reduce the false positive finding and to increase confidence in the results. In addition, the predicted INDELs/CNVs were compared with the previously identified and experimentally validated CNVs. The human genes and the biological pathways that are possibly affected by the predicted INDELs/CNVs were also identified.

3.1 INDELs and CNVs initially predicted by QuantiSNP and PennCNV

The computational prediction of INDELs/CNVs in this project was performed utilizing the QuantiSNP and PennCNV algorithms. As a result, in the entire cohort a total of 336,288 and 204,439 INDELs/CNVs were identified by QuantiSNP and PennCNV algorithms, respectively. **Table 3.1** summarizes the main features of these initially predicted INDELs/CNVs. In brief, QuantiSNP analysis yielded more INDEL/CNV predictions than the PennCNV analysis; the reason for this difference is not clear, however it is possibly due to the different HMMs and different parameters used by these algorithms [77,153,248]. Also, both algorithms predicted a higher portion of CNVs than INDELs. In addition, the number of deletions (either homozygous or heterozygous deletions) was higher than the number of duplications. In the case of duplications of two or more copies, QuantiSNP predicted a substantially higher number of variants compared to PennCNV (**Figure 3.1**).

Table 3.1: The main features of the INDELs and CNVs predicted by the QuantiSNP and PennCNV algorithms.

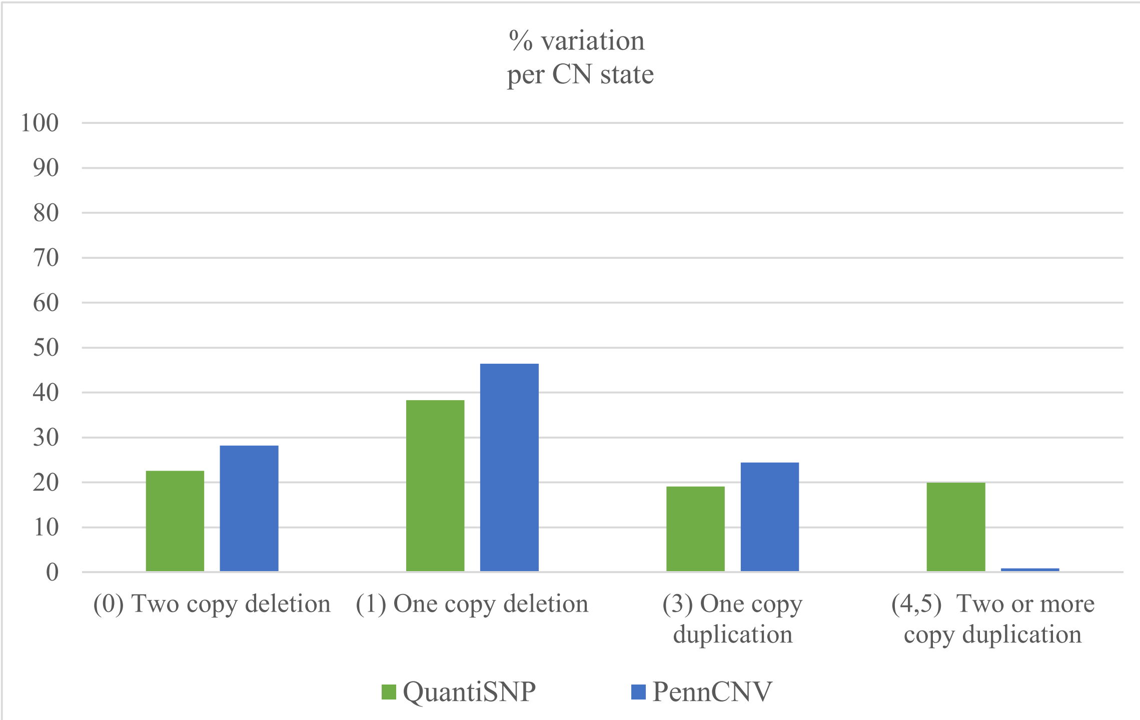
Number of INDELs/CNVs	**QuantiSNP**	**PennCNV**
Total predicted INDELs/CNVs in the cohort	336,288	204,439
Average number of INDELs/CNVs per individual	665.92	404.83

Type	**N**	**%**	**N**	**%**
INDELs	76,854	22.85	46,616	22.80
CNVs	259,434	77.15	157,823	77.20

INDELs/CNVs per CN state		**N**	**%**	**N**	**%**
(CN= 0)	Two copy deletion	76,035	22.61	57,698	28.22
(CN= 1)	One copy deletion	128,908	38.33	94,917	46.43
(CN= 3)	One copy duplication	64,217	19.1	49,983	24.45
*(CN= 4, 5)	Two or more copy duplication	67,128	19.96	1,841	0.90

N: Number, **CN**: Copy number state. *Please note that QuantiSNP assigns the CN state 4 for variants that exist in 4 copies and CN state 5 for variants that exist in 5 or more copies in a genome. However, PennCNV assigns the CN state 4 for variants that exist in 4 or more copies in a genom3942600/204439e.

Figure 3.1: Copy number states of the predicted INDELs and CNVs by QuantiSNP and PennCNV algorithms.



CN: Copy number state.

3.2 Post prediction QC analyses

Following the prediction of INDELs/CNVs, QC analyses were performed to exclude the low quality data from the QuantiSNP and PennCNV predictions (**Section 2.6.2**).

During the QC analysis of the initial QuantiSNP results, data of all patients included in this study fulfilled the QuantiSNP LRR_SD and BAF_SD criteria (**Appendix F. 1**); however, four patients were excluded because one patient had a CNV longer than 7.5 Mbp and three additional patients had excess INDEL/CNV calls (i.e. the predicted number of INDELs/CNVs for them exceeded the mean number of predicted variations +

3 SD) (**Appendix F. 2**). As a result, 501 patients fulfilled the QuantiSNP QC criteria. Additionally, a total of 250,819 initially predicted INDELs/CNVs (~ 74.5%) by the QuantiSNP algorithm were excluded because their size was < 10 bps, they contained binding sites for < 10 probes, or they had the maximum log Bayes factor (confidence score) < 30.

During the QC analysis of the initial PennCNV results, eight individuals failed to meet the QC criteria; four individuals had the LRR_SD > 0.28 (**Appendix F. 3**); one individual had a very long CNV (the same patient detected and excluded during the QC analysis of the QuantiSNP data); and three additional individuals had excessive INDEL/CNV calls (**Appendix F. 4**). Therefore, the data of 497 patients satisfied the QC criteria of PennCNV. In the case of variants predicted, a total of 45,389 INDELs/CNVs (~ 22.2%) predicted by PennCNV were < 10 bps, included binding sites for < 10 probes, or had a confidence score < 10. Hence, these INDEL/CNV predictions were considered to be of low quality and were excluded from the PennCNV data. **Appendix F. 5** shows the number of patients and summarizes the features of INDELs/CNVs that fulfilled the QuantiSNP and PennCNV QC criteria.

After the QC filtering, a total of 495 individuals passed the QC thresholds of both algorithms and constituted the final list of patients. The baseline features of these patients is presented in **Appendix F. 6**.

3.3 Additional filtering of the INDEL/CNV data

A series of inclusion/exclusion filtering was performed following the QC analysis to further reduce the methodological artifacts, to minimize the false positive findings, to eliminate the low quality data, and to exclude variants from the repetitive genomic regions (**Section 2.6.3**).

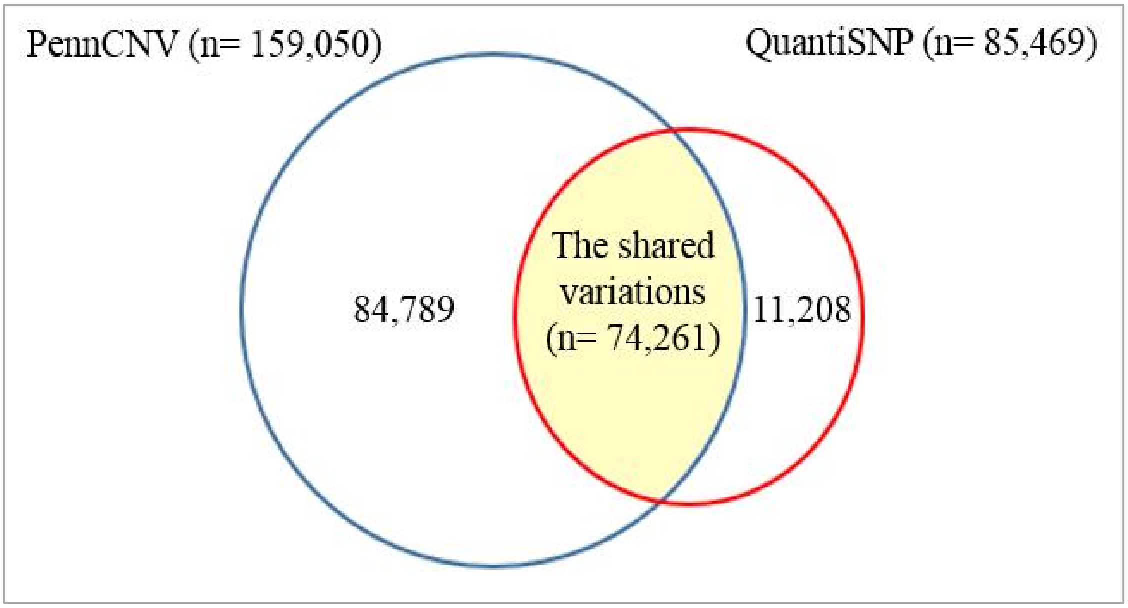
3.3.1 Overlaps between the INDELs/CNVs predicted by one algorithm in the same individual

Possible overlaps between the INDELs/CNVs predicted in the same patient were assessed in order to identify and exclude variants mistakenly predicted twice by either QuantiSNP or PennCNV (**Section 2.6.3.1**). As a result, no duplicated predictions were identified in the data set.

3.3.2 Overlaps between the variations predicted by both algorithms in the same individual

INDELs/CNVs that were predicted by both QuantiSNP and PennCNV algorithms, had the same CN state, and overlapped at least 50% of their length with each other were identified. As a result, 84,789 variations predicted by QuantiSNP and 11,208 variations predicted by PennCNV that did not have these characteristics were excluded (**Figure 3.2**). The remaining 74,261 variants in 495 patients and their main features are summarized in **Appendix G**. In summary, the number of CNVs were higher than INDELs; more than

Figure 3.2: Venn diagram showing the INDELs/CNVs predicted by QuantiSNP and PennCNV and the variations that are detected by both algorithms.



95% of the variants were homozygous or heterozygous deletions; and the variants with three or more copy numbers constituted only a small portion (~4.06%) of the variations.

Interestingly, a total of 62,567 of the variants (84.25%) predicted by both algorithms had identical start and send positions, suggesting a high-concordance between the results of PennCNV and QuantiSNP when a variant is detected by both of these algorithms.

3.3.3 Overlaps between the INDELs/CNVs and the highly repetitive DNA regions

Since highly repetitive DNA sequences, such as leukocyte Immunoglobulin-like receptor gene cluster and olfactory receptor gene sequences, as well as centromere and telomere regions, complicate the INDEL/CNV predictions (**Section 2.6.3.3**), variants that overlapped at least one bp with these DNA regions were excluded (n=2,905). As a result, 71,356 variations remained in the data set.

3.3.4 Overlaps between the predicted INDELs/CNVs and the previously identified CNVs

In this analysis (**Section 2.6.3.4**), the vast majority of the INDELs/CNVs predicted in this study (~97%; n=69,290) were detected to have at least 50% of their lengths overlapped with the variants that were identified by DNA analysis in three large-scale CNV studies [31,252,253]. These INDELs/CNVs were, therefore, highly likely to be existing in the DNAs of the patients, and hence constituted the final list of (high-confidence) variants in this study. In contrast, the remaining 3% of the predicted variations (n=2,066) were excluded from our final variations list, because there was a minimal evidence showing that they were not methodological artifacts. It is possible that some of these excluded variants may be in fact existing, but are rare or patient-specific variants; however this possibility can only be determined by DNA analysis.

The features of the high-confidence INDELs/CNVs in term of their length and CN state are summarized in **Table 3.2**. On average, 140 INDELs/CNVs are predicted per patient (**Figure 3.3**). Almost 80% of the high-confidence variants were CNVs and almost 98% of the variations were deletion variants (**Figure 3.4**). The number of variants between patient categories did not significantly differ, except for the age of onset (**Appendix H**).

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

		Number	
Number of patients		495	
Total INDELs/CNVs in the cohort		69,290	
Average number of INDELs/CNVs per individual		140	
Type		**N**	**%**
INDELs		14,364	20.73
CNVs		54,926	79.27
INDELs/CNVs per CN State		**N**	**%**
(CN= 0)	Two copy deletion	46,623	67.29
(CN= 1)	One copy deletion	21,144	30.51
(CN= 3)	One copy duplication	1,363	1.97
(CN= 4)	Two or more copy duplication	160	0.23

N: Number, **CN**: Copy number state.

Figure 3.3: Distribution of the number of predicted INDELs/CNVs in the patient cohort.

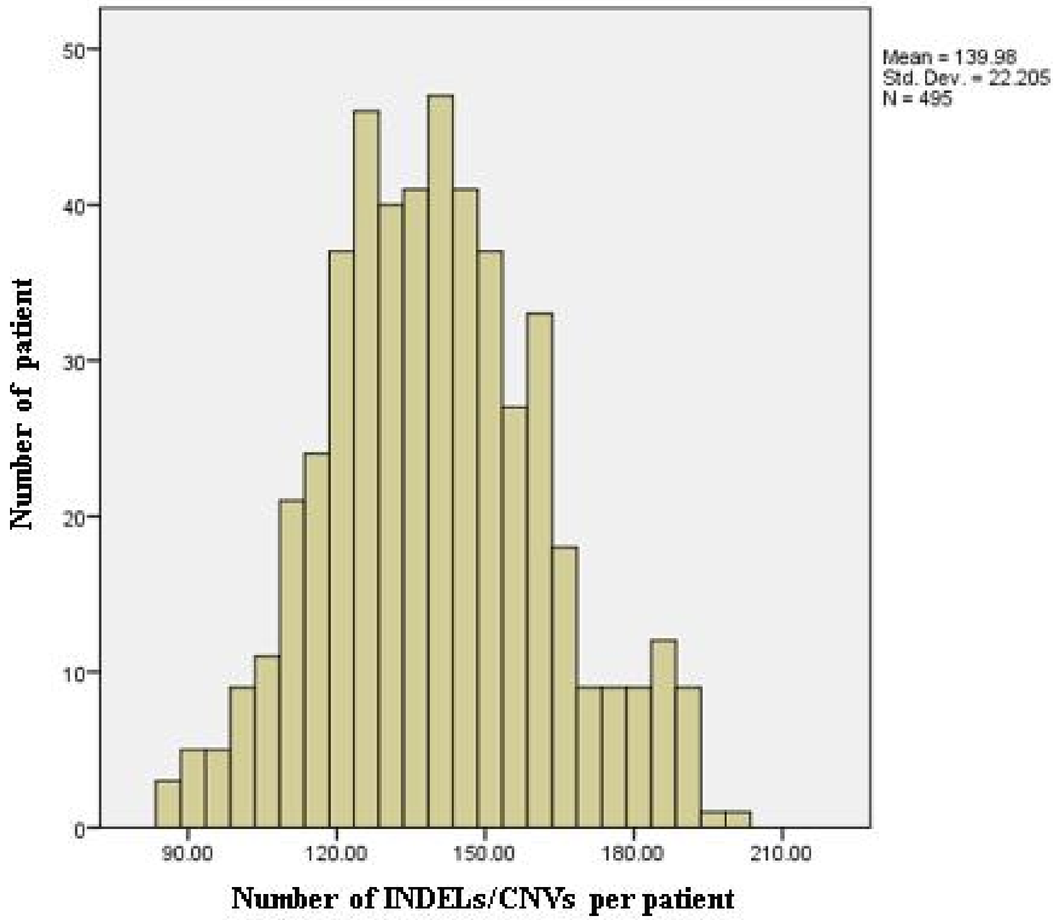
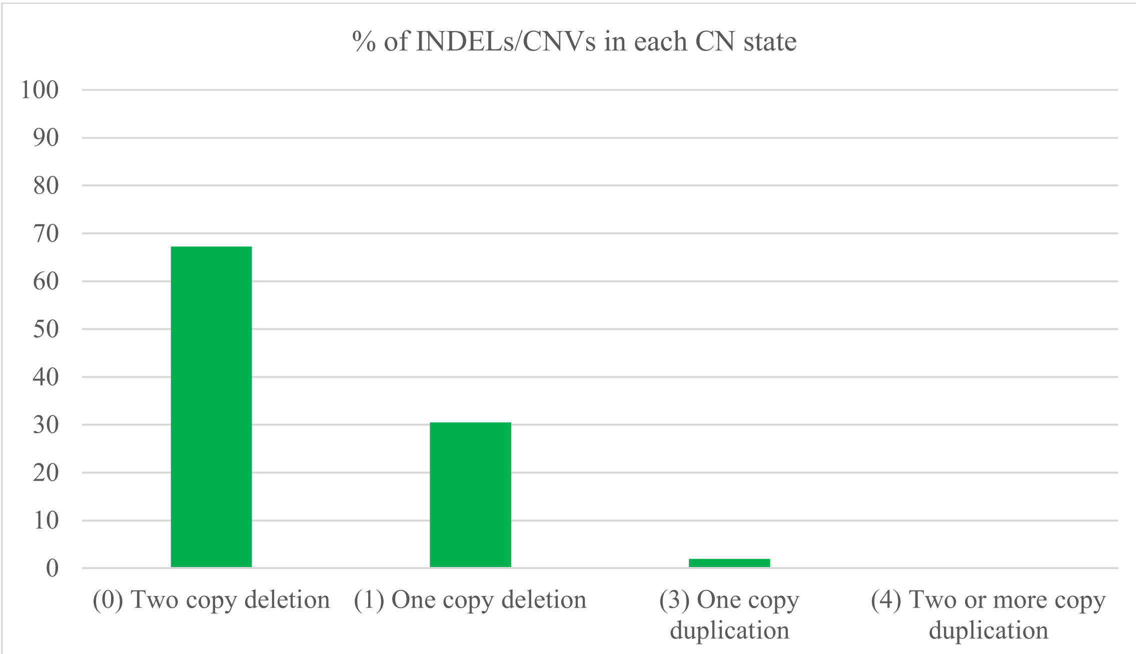


Figure 3.4: High-confidence INDELs/CNVs based on their copy number state.



3.4 The distinct, high-confidence INDELs/CNVs and the CNVRs

Once the high-confidence variants were identified, the next step was to identify the variants that had identical boundaries (i.e. identical start and end positions) (**Section 2.7**). In this thesis, these variants are referred to as "distinct, high-confidence variants". As also shown in **Table 3.3**, the high-confidence variants constituted 3,486 distinct INDELs/CNVs identified in at least one patient. The mean length of these distinct variations was ~35 kbps. CNVs made ~90% and INDELs formed the rest of the distinct INDELs/CNVs. Around 83% of the distinct INDELs/CNVs were rare variations that occurred in less than 5% of the individuals, whereas ~17% variations were common (frequency ≥ 5%).

Table 3.3: The main features of the distinct, high-confidence INDELs/CNVs identified in the study cohort.

Variable	**Number**
Total number of distinct INDELs/CNVs	3,486
Mean distinct INDEL/CNV length	35,187 bps

Length	**Number**	**%**
INDELs	360	10.33
CNVs	3,126	89.67

Frequency	**Number**	**%**
Rare INDELs/CNVs (< 5% of the patients)	2,891	82.93
Common INDELs/CNVs (≥ 5% of the patients)	595	17.07

***Number of INDELs/CNVs per CN state**	**Number**	**%**
INDELs/CNVs with two CN states	**2905**	**83.33**
(CN= 0) Two copy deletion	685	19.65
(CN= 1) One copy deletion	1,596	45.78
(CN= 3) One copy duplication	607	17.41
(CN= 4) Two or more copy duplication	17	0.49
INDELs/CNVs with multiple CN states	**581**	**16.67**
A. INDELs/CNVs with three CN states	**577**	**16.55**
CN= 0 or 1	543	15.58
CN= 0 or 3	7	0.20
CN= 0 or 4	2	0.06
CN= 1 or 3	13	0.37
CN= 3 or 4	12	0.34
B. Four INDELs/CNVs with four CN states	**4**	**0.12**
CN= 0, 3 or 4	1	0.03
CN= 0, 1 or 4	1	0.03
CN= 0, 1 or 3	2	0.06

CN: Copy number state. *the "normal" CN state of 2 copies is not shown.

Most of the distinct, high-confidence INDELs/CNVs (~81.01%) were deletions. Interestingly, around 0.75% of these variations were predicted as deletions in some patients and duplications in other patients. Taken together, 16.67% of these variations were multiallelic, while the majority of predicted variations in the study were biallelic, being either a deletion or duplication (**Table 3.3**).

Finally, these distinct, high-confidence INDELs/CNVs were clustered in 1,527 different CNVRs.

3.5 Genes that are possibly affected by the distinct, high-confidence INDELs/CNVs

The sequences of 2,209 INDELs/CNVs (63.4%) were identified to overlap with the sequences of 1,673 genes (genic INDELs/CNVs). These genes belonged to a variety of gene types (**Table 3.4**); the largest group of genes were the protein-coding genes (n=771, 46.1% combining the protein_coding genes, IG_D and IG_V genes in **Table 3.4**), followed by pseudogenes (n=422, 25.2%) and RNA-coding genes (n=339, 20.3%).

A significant portion of the variants (~42%, n=929) overlapped with the entire sequence of a gene, suggesting they may cause a gene dosage effect. Additionally, some genes (n=134) seem to be hot-spots for chromosomal rearrangements and CNVs as they were observed to have multiple INDELs/CNVs as shown in **Table 3.5**. The genes that are affected by multiple INDELs/CNVs contain cancer related genes, such as Complement Factor H-Related 2 (*CFHR2*); HEAT induced repeat containing (*HEATR4*); Protocadhein

Table 3.4: Classification of the genes that are likely to be affected by the INDELs/CNVs

Gene type	Gene type description from ENSEMBL	Number of affected genes
protein_coding gene	Genes and/or transcript that contains an open reading frame	748
pseudogene	Have homology to other proteins but generally the active homologous gene can be found at another locus	409
lincRNA	Long, intergenic non-coding RNA	229
rRNA	Non-coding RNAs predicted using sequences genes	7
snoRNA, snRNA	small nucleolar and small nuclear RNAs	43
miRNA	microRNA precursors	30
miscRNA	miscellaneous other RNA	30
Antisense	Genes or transcripts having transcripts that overlap the genomic regions (i.e. exon or introns) of a protein-coding locus on the opposite strand.	98
processed_transcript	Transcripts that do not contain an open reading frame	27
IG_D_gene and IG_V_gene	Gene that rearranges at the DNA level and codes the diversity (D, V) regions of the variable domain of immunoglobulins	23
IG_V_pseudogene	Locus that shares an evolutionary history with the Ig V gene but it has been mutated through frameshift and/or stop codon(s) that disrupt the open reading frame of immunoglobulins.	9
sense_intronic	Long non-coding transcript in introns of a coding gene that does not overlap with exons sequences.	8
sense_overlapping	Long non-coding transcript that contains a coding gene within one of its introns and on the same strand	8
polymorphic_pseudogenes	Pseudogene loci in one genome, but coding in other genomes.	4

This table is generated based on the gene type description in the ENSEMBL database [254].

Table 3.5: Genes possibly affected by the INDELs/CNVs.

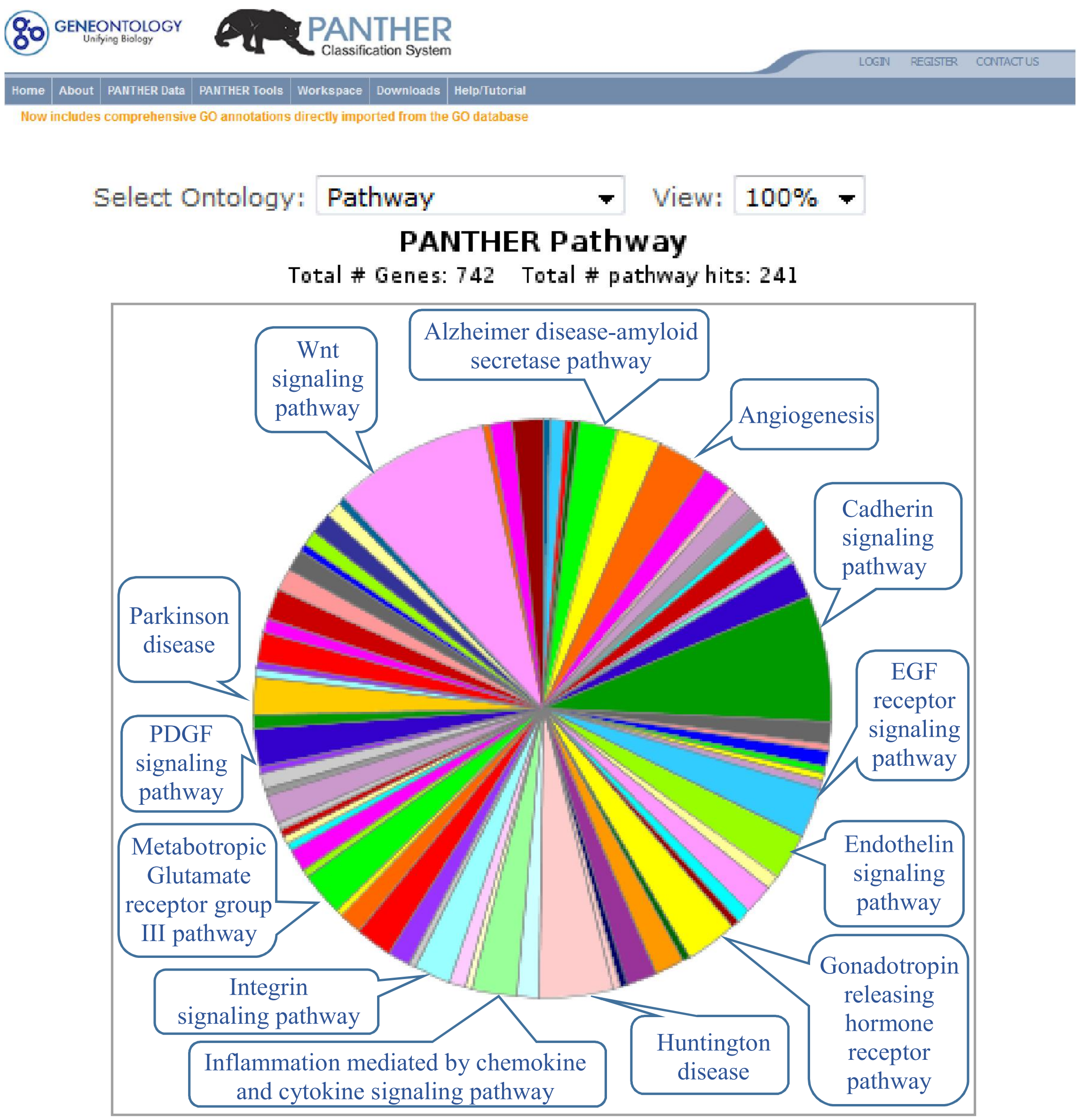
Affected genes	**Numbers**	**%**
Genes completely covered by INDELs/CNVs	659	39.39
Genes partially overlapped by INDELs/CNVs	880	52.60
Genes completely or partially overlapped by different INDELs/CNVs	134	8.01

Alpha 9 (*PCDHA9*); Major histocompatibility complex, class I, A (*HLA-A*); Tubulin, Alpha 8 (*TUBA8*); and Cytochrome P450, Family 2, Subfamily A, Polypeptide 7 (*CYP2A7*). The INDELs/CNVs that affect the genes in this group were identified as either rare or common variants (0.2-45.1% of the study patients), with a mean length of ~126 kbps. It is possible that these CNV hot-spot regions may contain sequences or signals that promote CNV formation, which is briefly explained in **Section 1.3**.

3.6 The biological pathways that may be affected by the distinct, high-confidence INDELs/CNVs

The PANTHER database [255,256] returned information for 742 out of the 1,673 genes overlapped with the distinct INDELs/CNVs. The results showed that these genes act in multiple biological pathways (n=241), including signaling, immune system, and neuro-hormone/neurotransmitter-related pathways. The largest group of genes was found to code for proteins functioning in the Wnt, cadherin, angiogenesis, integrin and chemokine/cytokine signaling pathways (**Figure 3.5**).

Figure 3.5: PANTHER database output showing the biological pathways possibly affected by the INDELs/CNVs.



Panther pathway	# genes	% of genes	gene symbol
Wnt signaling pathway (P00057)	21	2.8	PCDHA10, PCDH9, PCDHA4, PCDHA7, CDH19, CSNK1G2, PRKCB, PPP3CA, SMARCA2, PCDHA8, PCDHA5, PCDHA1, CTNNA3, PCDH15, PCDHA6, SMAD1,

				PCDHA3, PCDHA2, PCDHA9, CDH13, APC2
	Cadherin signaling pathway (P00012)	17	2.3	PCDHA10, PCDH9, PCDHA4, PCDHA7, NPL, CDH19, PCDHA8, PCDHA5, ERBB4, PCDHA1, CTNNA3, PCDH15, PCDHA6, PCDHA3, PCDHA2, PCDHA9, CDH13
	Huntington disease (P00029)	10	1.3	GRIK5, TP63, NPL, LARS2, C19orf25, GRIK4, PRODH, GRIK2, GRIN3A, RHOJ
	Angiogenesis (P00005)	7	0.9	ANGPT1, PDGFD, PIK3C2G, PRKCB, RBPJ, ARHGAP8, APC2
	EGF receptor signaling pathway (P00018)	7	0.9	PIK3C2G, PRKCB, ERBB4, MAPK10, NRG3, RHOJ, LRP5L
	Gonadotropin releasing hormone receptor pathway (P06664)	7	0.9	TGFBR3, BMP6, PRKCB, PPP3CA, SMAD1, LRP5L, CACNA1C,
	Alzheimer disease-presenilin pathway (P00004)	6	0.8	NPL, MLLT4, ERBB4, LRP1B, MMP23A, RBPJ
	Endothelin signaling pathway (P00019)	6	0.8	GUCY1A3, PIK3C2G, PRKCB, ADCY8, PRKG1, GUCY1A2
	Inflammation mediated by chemokine and cytokine signaling pathway (P00031)	6	0.8	NPL, CCL3L1, PIK3C2G, PRKCB, CCL4L1, PTEN
	Metabotropic glutamate receptor group III pathway (P00039)	6	0.8	GRIK5, GRIK4, SLC1A7, GRIK2, GRIN3A, WDR72
	Alzheimer disease-amyloid secretase pathway (P00003)	5	0.7	PRKCB,MAPK10, MAPK6, APBA1, CACNA1C
	CCKR signaling map (P06959)	5	0.7	PRKCB, PPP3CA, MAPK10, PRKG1, PTEN
	Integrin signalling pathway (P00034)	5	0.7	PIK3C2G, MAPK10, MAPK6, ABL1, RAP2A
	Ionotropic glutamate receptor pathway (P00037)	5	0.7	GRIK5, GRIK4, SLC1A7, GRIK2, GRIN3A
	PDGF signaling pathway (P00047)	5	0.7	FLI1, MAPK6, RPS6KA2, VAV3, ARHGAP8
	Parkinson disease (P00049)	5	0.7	SLC6A3, CSNK1G2, MAPK10, HSPA6, PARK2
	Apoptosis signaling pathway (P00006)	4	0.5	PRKCB, BOK, MAPK10, HSPA6
	B cell activation (P00010)	4	0.5	PRKCB, PPP3CA, MAPK10, VAV3

	FGF signaling pathway (P00021)	4	0.5	PIK3C2G, PRKCB,MAPK10, FGF12,
	Heterotrimeric G-protein signaling pathway-Gi alpha and Gs alpha mediated pathway (P00026)	4	0.5	CREB5, ADCY8, GNGT1, ADORA1
	Heterotrimeric G-protein signaling pathway-Gq alpha and Go alpha mediated pathway (P00027)	4	0.5	RASGRP3, PRKCB, GNGT1, ADORA1
	Nicotinic acetylcholine receptor signaling pathway (P00044)	4	0.5	NPL, MYO16, WDR72, CACNA1C,
	Ras Pathway (P04393)	4	0.5	MAPK10, EXOC2, C2orf73, RPS6KA2
	Synaptic_vesicle_trafficking (P05734)	4	0.5	UNC13C, SYT15, RIMS1, WDR72,
	p53 pathway (P00059)	4	0.5	BAI3, TP63, PIK3C2G, PTEN
	Axon guidance mediated by Slit/Robo (P00008)	3	0.4	SLIT2, ABL1, RHOJ
	Cytoskeletal regulation by Rho GTPase (P00016)	3	0.4	NPL, RHOJ, ARHGAP8
	Hypoxia response via HIF activation (P00030)	3	0.4	PIK3C2G, EGLN3, PTEN
	Interleukin signaling pathway (P00036)	3	0.4	IL4R, MAPK6, RPS6KA2
	Metabotropic glutamate receptor group I pathway (P00041)	3	0.4	GRIK5, PRKCB, GRIN3A
	Muscarinic acetylcholine receptor 1 and 3 signaling pathway (P00042)	3	0.4	PRKCB, GRIN3A, WDR72
	T cell activation (P00053)	3	0.4	PIK3C2G, PPP3CA, VAV3
	TGF-beta signaling pathway (P00052)	3	0.4	BMP6, MAPK10, SMAD1
	Ubiquitin proteasome pathway (P00060)	3	0.4	HACE1, ATG7, UBE2D3
	p53 pathway feedback loops 2 (P04398)	3	0.4	TP63, PIK3C2G, PTEN
	5HT2 type receptor mediated signaling pathway (P04374)	2	0.3	PRKCB, CACNA1C
	Axon guidance mediated by netrin (P00009)	2	0.3	PIK3C2G, UNC5C
	De novo purine biosynthesis (P02738)	2	0.3	AK4, NME7
	FAS signaling pathway (P00020)	2	0.3	IFLTD1, MAPK10
	GABA-B_receptor_II_signaling (P05731)	2	0.3	ADCY8, GABBR2

Insulin/IGF pathway-protein kinase B signaling cascade (P00033)	2	0.3	PIK3C2G, PTEN
Oxytocin receptor mediated signaling pathway (P04391)	2	0.3	PRKCB, CACNA1C
PI3 kinase pathway (P00048)	2	0.3	GNGT1, PTEN
Sulfate assimilation (P02778)	2	0.3	PAPSS1, REV1
Transcription regulation by bZIP transcription factor (P00055)	2	0.3	CREB5, GTF2A1L
VEGF signaling pathway (P00056)	2	0.3	PIK3C2G, PRKCB
5-Hydroxytryptamine biosynthesis (P04371)	1	0.1	TPH2
Adrenaline and noradrenaline biosynthesis (P00001)	1	0.1	SLC6A3
Allantoin degradation (P02725)	1	0.1	ALLC
Ascorbate degradation (P02729)	1	0.1	SHPK
Axon guidance mediated by semaphorins (P00007)	1	0.1	SEMA4D
Beta1 adrenergic receptor signaling pathway (P04377)	1	0.1	CACNA1C
Beta2 adrenergic receptor signaling pathway (P04378)	1	0.1	CACNA1C
DNA replication (P00017)	1	0.1	PRIM2
De novo pyrimidine deoxyribonucleotide biosynthesis (P02739)	1	0.1	NME7
De novo pyrmidine ribonucleotides biosythesis (P02740)	1	0.1	NME7
Dopamine receptor mediated signaling pathway (P05912)	1	0.1	SLC6A3
General transcription regulation (P00023)	1	0.1	GTF2A1L
Hedgehog signaling pathway (P00025)	1	0.1	UBR5
Heterotrimeric G-protein signaling pathway-rod outer segment phototransduction (P00028)	1	0.1	GNGT1
Histamine H1 receptor mediated signaling pathway (P04385)	1	0.1	PRKCB
Insulin/IGF pathway-mitogen activated protein kinase kinase/MAP kinase cascade (P00032)	1	0.1	RPS6KA2

Interferon-gamma signaling pathway (P00035)	1	0.1	MAPK10
Metabotropic glutamate receptor group II pathway (P00040)	1	0.1	WDR72
Methylmalonyl pathway (P02755)	1	0.1	PCCB
Muscarinic acetylcholine receptor 2 and 4 signaling pathway (P00043)	1	0.1	WDR72
N-acetylglucosamine metabolism (P02756)	1	0.1	NPL
Nicotine degradation (P05914)	1	0.1	CYP2A6
Nicotine pharmacodynamics pathway (P06587)	1	0.1	CACNA1C
Notch signaling pathway (P00045)	1	0.1	RBPJ
P53 pathway feedback loops 1 (P04392)	1	0.1	TP63
Pyridoxal phosphate salvage pathway (P02770)	1	0.1	PDXK
Pyruvate metabolism (P02772)	1	0.1	PC
Thyrotropin-releasing hormone receptor signaling pathway (P04394)	1	0.1	PRKCB
Vitamin B6 metabolism (P02787)	1	0.1	PDXK
p53 pathway by glucose deprivation (P04397)	1	0.1	TP63

Chapter 4: Discussion and conclusions

In the last decade, structural variations such as INDELs and CNVs, were discovered in the human genome and have been identified to contribute to human genetic variability, population diversity, and the susceptibility to diseases [28,49,110], including cancer [7,12]. The latest advances in high resolution SNP arrays and sophisticated calling algorithms have facilitated the genome-wide identification of INDELs/CNVs.

As part of a much larger project in the Savas lab, this research project aimed to computationally detect and characterize the INDEL/CNV profiles in the genomes of colorectal cancer patients and to identify the genes and biological pathways that may be affected by the predicted variations. The genome-wide signal intensity data used in this project was previously generated using the high resolution Illumina® Human Omni1_Quad_v1 SNP genotyping platform [139]. These data were used to computationally identify the INDEL/CNV profiles in the genomes of 505 unrelated Caucasian colorectal cancer patients recruited to the NFCCR[15,16,227]. The INDELs/CNVs were detected using two CNV calling algorithms, QuantiSNP and PennCNV [17,18]. To eliminate the low quality DNA samples and INDELs/CNVs in the results, stringent QC analyses and inclusion/exclusion filtering procedures were performed.

As summarized in **Figure 3.2**, the results of the QuantiSNP and PennCNV were notably different in term of the number of predicted INDELs/CNVs. This can be attributed to the different methodological bases of these algorithms, which was also observed in other studies [77]. However, when a variant was predicted by both algorithms (i.e. with at least 50% overlap), in 84.3% of the predictions the genomic positions (i.e.

boundaries) of the INDELs/CNVs predicted were identical (**Section 3.3.2**). This indicates a high-concordance rate for QuantiSNP and PennCNV when a variant is detected by both of them.

Subsequent to quality control analysis, ~97% of the detected INDELs/CNVs were found to be located in previously identified and experimentally validated CNVRs (**Section 3.3.4**). These INDELs/CNVs constituted the final list of 69,290 high-confidence variations identified in the genomes of 495 patients, with an average of 140 variants per patient. The average number of variations per individual in my results is similar to the findings of Horpaopan et al (2014) study (154 CNV events per patient), which also used the Illumina® Human Omni1_Quad_v1 SNP genotyping platform and utilized the QuantiSNP algorithm to identify the CNV in colorectal cancer patient from Germany [216].

The fact that nearly 97% of the INDELs/CNVs detected in this study were previously identified using DNA analyses in other individuals suggest that the variants detected by algorithms in this study are highly likely to exist in the patient DNAs (i.e. not likely to be methodological artifacts or false-positives). Furthermore, recent experiments performed by Ms. Georgia Skardasi in the Savas lab showed that the results of the computational analyses and the DNA analyses in 10 homozygously deleted CNVs were concordant in 93 – 100 % of the cases (*unpublished data*). These findings increase the confidence in the approach/quality control measures utilized (that aimed to reduce the false-positive predictions) and the results obtained during this project.

The high-confidence INDELs/CNVs detected in the patient genomes constituted 3,486 distinct INDELs/CNVs (**Section 3.4**). The sizes of these distinct, high-confidence variations ranged from 359 to 956,373 bps, with a mean length of ~35 kbps. Although both INDELs (sizes < 1kbp) and CNVs (sizes ≥ 1kbp) were detected in the examined genomes, CNVs constituted the largest portion of the identified variations (~90%). This is not surprising, because the QC approach utilized in this study was geared toward detecting CNVs rather than INDELs. For example, variants with sizes < 10 bps or detected by < 10 probes were eliminated from the results to increase the accuracy of our results. This inevitably resulted in excluding a portion of INDELs in this study.

As explained in **Section 1.4.2**, use of the signal intensity data obtained by SNP genotyping platforms leads to more robust prediction of deletions compared to duplications [134], and as expected around 81% of the distinct, high-confidence variants in this study were deletions. These results are in agreement with the findings of several CNV studies that used a similar approach [77,134,153]. Additionally, ~83% of these distinct, high-confidence INDELs/CNVs were rare, occurring in less than 5% of the patients, while ~17% of them were common detected in at least 5% of the study population. These variants may be an interest for further studies related to human diseases and population diversity [7]. For example, rare germline CNVs may lead/contribute to high-penetrant genetic disorders including hereditary colon cancer syndromes such as FCCX. Also, common germline CNVs may have roles as low penetrant alleles in cancer predisposition, progression, or survival outcomes. Thus, both the rare and common INDELs/CNVs that are identified in this study constitute an interesting set of variants, especially in colorectal cancer.

As a matter of fact, some of the rare and common CNVs were previously linked to colorectal cancer. The first example is a deletion CNV at 2p22.3 (Chr2:34,698,447-34,736,476) that was identified in ~37% (182/495) of our study patients. This variant overlaps with three previously identified CNVRs (dgv691e199, esv26780, and nsv514059 [83]) and with a deletion CNV (Chr2:34,699,314-34,737,163) that was reported by Fernandez et al (2014) as associated with the susceptibility to colorectal cancer [82]. The second example is a rare CNV consisting of 108,251 bps at 15q14 (Chr15:34,730,758-34,851,798). This CNV overlaps with a known CNVR (esv27955 [83]) as well as deletion CNV at (Chr15:34,700,683-34,830,944) that associates with the colorectal cancer risk [82]. There are other variants that are predicted in this study and linked to colorectal cancer by previous studies, which are briefly discussed in the following section.

My results also showed that 63.4% of the distinct, high-confidence INDELs/CNVs partially or completely overlapped with genes (n = 1,673). About 66% of the rare variations (1,926/2,891) were identified to overlap with genes (n = 1,538). In contrast, only 47.6% of the common INDELs/CNVs (283/595) were detected to overlap with genes (n = 135). This difference between the rare and common genic INDELs/CNVs may be, at least partially, due to the negative selection acting on the variants that affect the biologically critical genes and that are therefore kept at low frequencies in populations [138].

Table 4.1 shows examples of rare and common genic INDELs/CNVs that were identified to possibly affect genes. Interestingly, some of these variants overlap with cancer-related genes. For example, one of these genic CNVs is a 10,894 bps heterozygous deletion CNV on 1q44 (Chr1:245,636,915-245,647,809) detected in ~6.06% (30/495) of the patients (**Table 4.1**). This CNV locates in previously identified CNVRs (esv2673051 and esv22570 [83]) and deletes the exon 10 of the Kinesin Family Member 26B (*KIF26B*) gene. KIF26B protein plays a significant role during

Table 4.1: Examples of genic CNVs that were identified to affect genes.

genic CNVs						Genes		
	Chr	**Start position**	**End position**	**CN state**	**Frequency (%)**	**Gene Symbol**	***Gene Function**	****Location of the CNV within the gene**
Common CNVs	1	152,556,085	152,586,939	0 2 3	33.13 65.25 1.62	*LCE3B, LCE3C*	Precursors of the cornified envelope of the stratum corneum	Covers the whole gene
	1	169,207,360	169,241,309	0 1 2	11.72 33.33 54.95	*NME7*	Candidate oncogene in neuroblastoma	Located within the intron 7 of the gene
	1	245,636,915	245,647,809	1 2	6.06 93.94	*KIF26B*	Play a role in the regulation of cell-cell adhesion, embryogenesis and kidney development	Spans over introns 10 and 11, and deletes exon 10 of the gene
	2	54,565,729	54,567,590	0 1 2	28.08 0.20 71.72	*C2orf73*	Signal transducer and activator of transcription	Located within the intron 1 of the gene
	2	100,103,752	100,105,013	0 2	20 80	*REV1*	Deoxycytidyl transferase (involved in DNA repair)	Located within the intron 1 of the gene
	2	213,370,272	213,370,806	0 1 2	2.22 18.99 78.79	*ERBB4*	Tyrosine-protein kinase, epidermal growth factor receptor, cell to cell adhesion	Located within the intron 2 of the gene
	3	99,628,822	99,629,567	0 1 2	10.30 15.76 73.94	*FILIP1L*	Regulator of the antiangiogenic activity on endothelial cells	Located within the intron 4 of the gene

Common CNVs	4	107,058,139	107,063,124	0 1 2	6.06 26.26 67.68	*TBCK*	Regulation of molecular function, cell growth and cell proliferation	Located within the intron 24 of the gene
	4	162,449,613	162,451,188	0 1 2	3.23 5.86 90.91	*FSTL5*	Role in calcium ion binding and localize in cytoplasm, extracellular space and extracellular region.	Located within the intron 10 of the gene
	4	186,441,932	186,444,110	0 2	20.40 79.60	*PDLIM3*	Transcription factor, plays a role in organization of actin filament arrays within muscle cells	Located within the intron 3 of the gene
	16	78,373,700	78,384,735	0 1 2	24.24 6.66 69.09	*WWOX*	Tumor suppressor, gene plays a role in apoptosis	Located within the intron 5 of the gene
	17	724,239	724,598	0 2	34.55 65.45	*NXN*	Redox-dependent negative regulator of the Wnt signaling pathway, transcriptional regulator	Located within the intron 4 of the gene
	17	34,597,211	34,645,966	1 2	9.09 90.91	*CCL3L1*	Chemotactic for lymphocytes and monocytes	Cover the whole gene
	17	55,688,120	55,689,796	0 2	35.96 64.04	*MSI2*	RNA binding protein that regulates the expression of target mRNAs at the translation level	Located within the intron 8 of the gene

Rare CNVs	8	4,123,211	4,124,156	0 2	1.01 98.99	*CSMD1*	Tumor suppressor gene	Located within the intron 3 of the gene
	8	15,402,936	15,410,250	1 2	0.20 99.80	*TUSC3*	Tumor suppressor gene; may be involved in N-glycosylation	Located within the intron 1 of the gene
	20	41,178,700	41,243,236	1 2	0.20 99.8	*PTPRT*	Plays a role in cell growth, differentiation, mitotic cycle, and oncogenic transformation	Located within the intron 7 of the gene
	20	39,882,637	39,889,183	1 2	0.60 99.40	*ZHX3*	Acts as a transcriptional repressor	Located within the intron 2 of the gene

Chr: Chromosome number, **CN**: Copy number state. *Information obtained from the GeneCards encyclopedia [257]. **Based on the information in the Ensembl database [254].

embryogenesis, mainly limb and kidney development [258]. It has also been noted that the change in the expression of this gene is associated with the development and progression of various human cancers. For example, the upregulation of the *KIF26B* gene has been identified to be significantly associated with the risk [259] and short survival [260] of breast cancer patients. Additionally, this CNV is one of the variants that may be related to colorectal cancer; a study performed by Horpaopan et al (2014) identified a deletion CNV in the *KIF26B* gene in colorectal cancer patients from Germany and suggested an important role for the *KIF26B* gene in colorectal cancer development [216]. Therefore the CNV identified in this study may affect the expression of the *KIF26B* gene and contribute to the risk or progression of colorectal cancer.

Another 33,949 bps long deletion CNV (located on Chr1:169,207,360-169,241,309) was found as a common variation in ~45% (223/495) of the patients either as a homozygous or heterozygous deletion (**Table 4.1**). This CNV overlaps with a previously identified CNVR reported in the DGV (esv22143 [83]) and locates in the intron 7 of another cancer-related gene, the nonmetastatic cells 7 (*NME7*) gene. This gene codes for the NME7 enzyme acting as a nucleoside-diphosphate and protein histidine kinase. This protein is a component of γ-tubulin ring complex (γTuRC) that facilitates microtubule nucleation of the γTuRC during the cell cycle [261]. *NME7* gene has been reported by Diskin et al (2010) as a predisposition gene and a candidate oncogene in neuroblastoma, and a larger deleted CNVR (61,007 bps) including the region of this CNV, has been reported to be associated with the risk of neuroblastoma [262].

Another example of common genic CNVs includes a deletion CNV at 4q32.2 (Chr4: 162,449,613-162,451,188) that spans a genomic sequence of 1,575 bps (**Table**

4.1). This CNV was detected as homozygous or heterozygous deletion in ~9.09% (45/495) of the patient cohort. This CNV overlaps with three known CNVRs (dgv987e199, esv2666113 and esv29667 [83]) and is located in the intron 10 of the Follistatin-Like 5 gene (*FSTL5*). *FSTL5* encodes an extracellular matrix protein that is critical for normal physiological function [263]. Even though the role of this gene in cancer is not well understood, it has been reported to be related to known etiologic pathways involved in colorectal cancer development [264], such as the Transforming Growth Factor β (TGF-β) signaling pathway [265,266]. Therefore, this CNV is another interesting genetic variation to investigate in relation to colorectal cancer.

Additionally, some of the INDELs/CNVs identified in this project were found to overlap with tumor suppressor genes. For instance, a 5,345 bps deletion CNV on 16q23.1 (Chr16:78,373,700-78,384,735) has been identified in ~ 31% (153/495) of the patients as a homozygous or heterozygous deletion (**Table 4.1**). This CNV overlaps with three previously identified CNVRs according to the DGV database (dgv508e199, esv22307 and nsv514817 [83]). This deletion CNV eliminates a sequence from the intron five of the WW domain containing oxidoreductase (*WWOX*) gene. *WWOX* encodes a protein with two WW domains and a short-chain dehydrogenase (adh) domain that plays an important role in apoptosis in mouse and human [267]. Moreover, multiple studies revealed a significant association between a deleted region in the *WWOX* gene and tumor development, suggesting that the *WWOX* gene acts as a tumor suppressor gene [267]. Further studies can reveal whether this common CNV affects the expression levels of the *WWOX* gene. Overall, it is possible that these (**Table 4.1**) and other common INDELs/CNVs may act as low penetrant or modifying alleles and thus can be prioritized for further research in relation to in colorectal cancer susceptibility or outcome.

As also mentioned before, the majority of the variants detected were rare, with allele frequencies less than 5%. Examples of rare variations include a 945 bps deletion INDEL at 8p23.2 (Chr8:4,123,211-4,124,156), which was found in ~1% (5/495) of the study patients and locates in two previously identified CNVRs (esv2670772 and esv26084 [83]) (**Table 4.1**). This INDEL overlaps with intron 3 of the CUB and sushi domain-containing protein 1 (*CSMD1*) gene. *CSMD1* is a tumor suppressor gene that codes for transmembrane protein having a role in cell adhesion and cancer [268]. In a recent study comparing the tumor and non-tumor tissues, this gene was reported to be frequently (~17%) deleted in colorectal tumors, further supporting its role as a tumor suppressor gene [269]. This CNV is therefore an exciting candidate for further studies in colorectal cancer.

The last variant that I will discuss and summarized in **Table 4.1** is another rare (~0.2%) genic variant detected in our patient cohort. This variant is a 7,314 bps deletion CNV at 8p22 (Chr8: 15,402,936-15,410,250) located within two previously identified CNVRs (esv28500 and nsv514477 [83]). This CNV deletes a part of intron one of the tumor suppressor candidate 3 (*TUSC3*) gene, which is involved in N-glycosylation [257] and has a possible role in malignant ovarian tumors [223]. Interestingly, the CNV identified in this study falls within a larger CNV that has been previously reported to be significantly associated with colorectal cancer susceptibility and progression [269]. Similar to other CNVs discussed, whether or not this CNV contributes to colorectal cancer risk or progression therefore warrants further investigation.

I also examined the types of genes and biological pathways that may be affected by the predicted INDELs/CNVs (**Section 3.5** and **Section 3.6**). As shown in **Table 3.4**, in this study protein coding genes constituted the largest group (n=771) of the genes that are likely to be affected by the INDELs/CNVs. The biological pathway analysis using the PANTHER database returned information for (742/1,673) of the possibly affected genes. Interestingly, 709 of these PANTHER gene hits were protein coding genes. This indicates a possible bias in the PANTHER database towards protein coding genes [255]. According to PANTHER database, 742 genes that overlapped with the INDELs/CNVs were identified to act in several biological pathways, including signaling, immune system, and neurohormone/neurotransmitter related pathways. In particular, the largest group of these genes code for proteins acting in the Wnt, cadherin, integrin, and angiogenesis pathways (**Figure 3.5**). This is interesting as these pathways are also cancer-related pathways. For example the role of Wnt pathway in colorectal cancer [270,271] and the role of angiogenesis in cancer progression [272,273] are well known. These results suggest that at least some of the INDELs/CNVs predicted in this study may be potentially important in modifying the carcinogenesis-related biological processes in colorectal cancer.

In addition to the variants overlapping with protein coding genes, nearly 20% of the genic INDELs/CNVs in this study partially or entirely overlap with the RNA-coding genes, particularly long intergenic non-coding RNAs and microRNAs (**Table 3.4**). These RNA species are known to play regulatory roles in gene expression [274] and several recent studies emphasized their possible roles in cancer in addition to their normal physiological functions [275]. It is hence possible that these INDELs/CNVs may disturb or alter the expression levels (either increase or decrease depending on the copy number) of such RNA genes in individuals. Thus, it is reasonable to hypothesize that such alterations could possibly have a biological role in colorectal cancer, as abnormalities in non-coding RNA levels could potentially contribute to abnormalities in the expression levels of other genes.

So far, a limited number of other studies have been performed to detect the germline CNVs in the genomes of Caucasian colorectal cancer patients, and these studies examined patients with hereditary or familial colorectal cancer only. For example, Venkatachalam et al (2010) detected the genome-wide germline CNVs in 41 early-onset familial colorectal cancer patients [218], while Horpaopan et al (2015) identified the genome-wide germline CNVs in 221 *APC* and *MUTYH*-mutation negative patients using the QuantiSNP algorithm and investigated the roles of rare CNVs in FAP [216]. Another study published by Masson et al (2013) investigated the contribution of CNVs to disease risk in HNPCC patients [276]. Since our cohort mostly (~94%) consists of sporadic colorectal cancer patients, I believe my results not only substantially expand the current limited body of knowledge related to CNVs, but will also be useful in examination and understanding of particularly sporadic colorectal cancer patients.

Similar to other studies using a genome-wide CNV identification approach, the present study has some strengths and limitations. The strengths of the study are as follows; a) to our knowledge, this is one of the first comprehensive genome-wide studies that computationally identified the germline INDEL/CNV profiles in colorectal cancer patients, especially sporadic colorectal cancer patients. This study, therefore, created new information that will be indispensable for further studies in colorectal cancer; b) the

Illumina® Human Omni1 QuadV1 platform used in this analysis is considered to be one of the high-resolution genome-wide platforms having a high number (more than a million) of SNP and CNV probes. This high marker density increases the chances of accurate CNV detection and better estimation of variant borders; c) as recommended by most of the recent CNV studies, two CNV calling algorithms were used in this study to improve the accuracy of INDEL/CNV prediction, which decreases the false-positive prediction rate; and d) around 97% of the INDELs/CNVs were detected to be located in the previously identified and experimentally validated CNVRs. Together with the unpublished DNA analysis results obtained in the Savas lab, these findings increase my confidence in the approach I followed during this study.

The approach used to detect INDELs/CNVs in this study also has some limitations; a) even though the Illumina® Human Omni1 QuadV1 platform is a high resolution platform, its genome coverage is reported to be 93% for the CEU (i.e. Caucasian) population [137], therefore variants from a portion of the genome remains unidentified in this study; b) because of the complications in analysis of data from the sex chromosomes, INDELs/CNVs were predicted only from the autosomal chromosomes; c) although stringent QC and inclusion/exclusion analyses were done to minimize the false positive findings, these QC criteria might also increase the false negative prediction and hence possibly eliminated true INDELs and CNVs from my results; d) since this study primarily aimed to detect CNVs (sizes ≥ 1kbp) , a significant portion of the INDELs in the patient genomes were possibly eliminated during the QC analyses; e) the signal-intensity based approach used to identify the INDELs/CNVs in this project tends to under-detect duplications and thus many duplication variants possibly remained undetermined; and f) the data presented was obtained from Caucasian patients and thus may not be fully relevant for studies investigating other ethnicities.

In conclusion, this study is one of the first genome-wide studies that computationally identified and characterized the germline INDELs/CNVs in the genomes of Caucasian colorectal cancer patients. My results point to a number of INDELs and CNVs that may be potentially important in the susceptibility or prognosis of colorectal cancer. Further studies in the Savas lab will investigate these variants in detail including whether they are associated with the disease characteristics and survival outcomes in colorectal cancer. Overall, the results of this study add to the growing body of scientific knowledge on INDELs/CNVs and are expected to expedite further research and discoveries in colorectal cancer.

REFERENCES

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

2. Herszenyi L, Tulassay Z. Epidemiology of gastrointestinal and liver tumors. *Eur Rev Med Pharmacol Sci*. 2010;14(4):249-258.

3. Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian cancer statistics 2014. 2014;ISSN 0835-2976(10/20).

4. Glessner JT, Wang K, Cai G, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*. 2009;459(7246):569-573.

5. Rovelet-Lecrux A, Hannequin D, Raux G, et al. APP locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. *Nat Genet*. 2006;38(1):24-26.

6. Singleton AB, Farrer M, Johnson J, et al. Alpha-synuclein locus triplication causes parkinson's disease. *Science*. 2003;302(5646):841.

7. Shlien A, Malkin D. Copy number variations and cancer. *Genome Med*. 2009;1(6):62.

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

9. Liu W, Sun J, Li G, et al. Association of a germ-line copy number variation at 2p24.3 and risk for aggressive prostate cancer. *Cancer Res*. 2009;69(6):2176-2179.

10. Cybulski C, Wokołorczyk D, Huzarski T, et al. A deletion in CHEK2 of 5,395 bp predisposes to breast cancer in poland. *Breast Cancer Res Treat*. 2007;102(1):119-122.

11. Walsh T, Casadei S, Coats KH, et al. Spectrum of mutations in BRCA1, BRCA2, CHEK2, and TP53 in families at high risk of breast cancer. *JAMA*. 2006;295(12):1379-1388.

12. Kuiper RP, Ligtenberg MJ, Hoogerbrugge N, Geurts van Kessel A. Germline copy number variation and cancer risk. *Curr Opin Genet Dev*. 2010;20(3):282-289.

13. Zogopoulos G, Ha KC, Naqib F, et al. Germ-line DNA copy number variation frequencies in a large north american population. *Hum Genet*. 2007;122(3-4):345-353.

14. Ligtenberg MJ, Kuiper RP, Chan TL, et al. Heritable somatic methylation and inactivation of MSH2 in families with lynch syndrome due to deletion of the 3′ exons of TACSTD1. *Nat Genet*. 2008;41(1):112-117.

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

16. Xu W, Xu J, Shestopaloff K, et al. A genome wide association study on Newfoundland colorectal cancer patients' survival outcomes. *Biomark Res*. 2015;3:6.

17. Colella S, Yau C, Taylor JM, et al. QuantiSNP: An objective bayes hidden-markov model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*. 2007;35(6):2013-2025.

18. Wang K, Li M, Hadley D, et al. PennCNV: An integrated hidden markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res*. 2007;17(11):1665-1674.

19. Galton F. The history of twins, as a criterion of the relative powers of nature and nurture. *Journal of the Anthropological Institute of Great Britain and Ireland*. 1876:391-406.

20. Watson JD, Crick FH. Genetical implications of the structure of deoxyribonucleic acid. *JAMA*. 1993;269(15):1967-1969.

21. Freeman S. Biological science. In: 3rd ed. San Francisco, USA: Pearson Benjamin Cummings; 2008:430-453.

22. Weaver R&, Philip. Genetics. In: Kane K, 2nd ed. USA: Wm. C. Brown; 1989:282-290.

23. Britannica EB. "Germinal mutation." *Encyclopaedia britannica. encyclopaedia britannica online academic edition*. encyclopædia britannica inc. http://www.britannica.com/EBchecked/topic/231781/germinal-mutation. Updated 2014. Accessed 07/22, 2014.

24. Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. *Nature*. 2008;453(7191):56-64.

25. Wang K, Chen Z, Tadesse MG, et al. Modeling genetic inheritance of copy number variations. *Nucleic Acids Res*. 2008;36(21):e138.

26. Human Genome Structural Variation Working Group, Eichler EE, Nickerson DA, et al. Completing the map of human genetic variation. *Nature*. 2007;447(7141):161-165.

27. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet*. 2009;10(4):241-251.

28. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. *Science*. 2004;305(5683):525-528.

29. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 1998;280(5366):1077-1082.

30. Kwok P. Single nucleotide polymorphisms. In: Pui-Yan Kwok, ed. *Single nucleotide polymorphisms, methods and protocols.* 2nd ed. Springer; 2003:1-12.

31. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*. 2010;467(7319):1061-1073.

32. Cargill M, Altshuler D, Ireland J, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet*. 1999;22(3):231-238.

33. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.

34. Ecker JR, Bickmore WA, Barroso I, Pritchard JK, Gilad Y, Segal E. Genomics: ENCODE explained. *Nature*. 2012;489(7414):52-55.

35. Gibbs RA, Belmont JW, Hardenbol P, et al. The international HapMap project. *Nature*. 2003;426(6968):789-796.

36. Yang J, Manolio TA, Pasquale LR, et al. Genome partitioning of genetic variation for complex traits using common SNPs. *Nat Genet*. 2011;43(6):519-525.

37. Sanseau P, Agarwal P, Barnes MR, et al. Use of genome-wide association studies for drug repositioning. *Nat Biotechnol*. 2012;30(4):317-320.

38. Newton-Cheh C, Johnson T, Gateva V, et al. Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet*. 2009;41(6):666-676.

39. Program CAM, Ober C, Nicolae DL, Mexico City Childhood Asthma Study (MCAAS). Meta-analysis of genome-wide association studies of asthma in ethnically diverse north american populations. *Nat Genet*. 2011;43(9):887-892.

40. Zeggini E, Scott LJ, Saxena R, et al. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet*. 2008;40(5):638-645.

41. Cho YS, Chen C, Hu C, et al. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east asians. *Nat Genet*. 2012;44(1):67-72.

42. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. *Nat Genet*. 2007;39(11):1315-1317.

43. Lu S, Pardini B, Cheng B, et al. Single nucleotide polymorphisms within interferon signaling pathway genes are associated with colorectal cancer susceptibility and survival. *PloS One*. 2014;9(10):e111061.

44. De Gobbi M, Viprakasit V, Hughes JR, et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science*. 2006;312(5777):1215-1217.

45. Li G, Pan T, Guo D, Li LC. Regulatory variants and disease: The E-cadherin -160C/A SNP as an example. *Mol Biol Int*. 2014;2014: (Article ID 967565).

46. Litt M, and Luty JA. A hypervariable microsatellite revealed by in vitro amplification of dinucleotide repeat within the cardiac muscle action gene. *Am. J. Hum. Genet*. 1989;4:397-401.

47. Weber JL. Human DNA polymorphisms based on length variations in simple sequence tandem repeats. in "genome analysis series" (S.Tilghman and K. Davies, eds.). *Genetic and Physical Mapping, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.* 1990;1:159-181.

48. Butler JM. Short tandem repeat typing technologies used in human identity testing. *BioTechniques*. 2007;43(4):2-5.

49. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: Changing the landscape of chromosomes and design of disease studies. *Hum Mol Genet*. 2006;15 Spec No 1:R57-66.

50. Ellegren H. Microsatellites: Simple sequences with complex evolution. *Nat Rev Genet*. 2004;5(6):435-445.

51. Kimmel M, Chakraborty R, King JP, Bamshad M, Watkins WS, Jorde LB. Signatures of population expansion in microsatellite repeat data. *Genetics*. 1998;148(4):1921-1930.

52. Walker FO. Huntington's disease. *The Lancet*. 2007;369(9557):218-228.

53. Santoro MR, Bray SM, Warren ST. Molecular mechanisms of fragile X syndrome: A twenty-year perspective. *Annu Rev Pathol: Mechanisms of Disease*. 2012;7:219-245.

54. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. *Nat Genet*. 2004;36(9):949-951.

55. Mills RE, Luttig CT, Larkins CE, et al. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Res*. 2006;16(9):1182-1190.

56. Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437-455.

57. Levy S, Sutton G, Ng PC, et al. The diploid genome sequence of an individual human. *PLoS Biology*. 2007;5(10):e254.

58. Wheeler DA, Srinivasan M, Egholm M, et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature*. 2008;452(7189):872-876.

59. Mills RE, Pittard WS, Mullaney JM, et al. Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res*. 2011;21(6):830-839.

60. Jiang Y, Turinsky AL, Brudno M. The missing indels: An estimate of indel variation in a human genome and analysis of factors that impede detection. *Nucleic Acids Res*. 2015;43(15):7217-7228.

61. Mullaney JM, Mills RE, Pittard WS, Devine SE. Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet*. 2010;19(R2):R131-6.

62. Hu J, Ng PC. Predicting the effects of frameshifting indels. *Genome Biol*. 2012;13(2):R9.

63. Ley TJ, Mardis ER, Ding L, et al. DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*. 2008;456(7218):66-72.

64. Chen K, McLellan MD, Ding L, et al. PolyScan: An automatic indel and SNP detection approach to the analysis of human resequencing data. *Genome Res*. 2007;17(5):659-666.

65. Andersen V, Christensen J, Overvad K, Tjonneland A, Vogel U. Polymorphisms in NFkB, PXR, LXR and risk of colorectal cancer in a prospective study of danes. *BMC Cancer*. 2010;10:484.

66. Almal SH, Padh H. Implications of gene copy-number variation in health and diseases. *J Hum Genet*. 2011;57(1):6-13.

67. Korbel JO, Urban AE, Grubert F, et al. Systematic prediction and validation of breakpoints associated with copy-number variants in the human genome. *Proc Natl Acad Sci U S A*. 2007;104(24):10110-10115.

68. McCarroll SA, Hadnott TN, Perry GH, et al. Common deletion polymorphisms in the human genome. *Nat Genet*. 2005;38(1):86-92.

69. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet*. 2008;40(10):1253-1260.

70. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*. 2008;40(10):1166-1174.

71. Rabbitts T. Chromosomal translocations in human cancer. *Nature*. 1994;372:141-149.

72. Freeman JL, Perry GH, Feuk L, et al. Copy number variation: New insights in genome diversity. *Genome Res*. 2006;16(8):949-961.

73. Chang JC, Kan YW. Beta 0 thalassemia, a nonsense mutation in man. *Proc Natl Acad Sci U S A*. 1979;76(6):2886-2889.

74. Bowden DK, Hill AV, Higgs DR, Oppenheimer SJ, Weatherall DJ, Clegg JB. Different hematologic phenotypes are associated with the leftward (-alpha 4.2) and rightward (-alpha 3.7) alpha+-thalassemia deletions. *J Clin Invest*. 1987;79(1):39-43.

75. Driscoll DA, Spinner NB, Budarf ML, et al. Deletions and microdeletions of 22q11. 2 in velo-cardio-facial syndrome. *Am J Med Genet*. 1992;44(2):261-268.

76. Lejeune J, Turpin R, Gautier M. Chromosomic diagnosis of mongolism. *Arch Fr Pediatr*. 1959;16:962-963.

77. Pinto D, Darvishi K, Shi X, et al. Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*. 2011;29(6):512-520.

78. Tsuang DW, Millard SP, Ely B, et al. The effect of algorithms on copy number variant detection. *PLoS One*. 2010;5(12):e14456.

79. Perry GH, Ben-Dor A, Tsalenko A, et al. The fine-scale and complex architecture of human copy-number variation. *AJHG*. 2008;82(3):685-695.

80. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet*. 2008;40(10):1199-1203.

81. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. *Nat Genet*. 2007;39:S37-S42.

82. Fernandez-Rozadilla C, Cazier J, Tomlinson I, et al. A genome-wide association study on copy-number variation identifies a 11q11 loss as a candidate susceptibility variant for colorectal cancer. *Hum Genet*. 2014;133(5):525-534.

83. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The database of genomic variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014;42(Database issue):D986-92.

84. Korbel JO, Urban AE, Affourtit JP, et al. Paired-end mapping reveals extensive structural variation in the human genome. *Science*. 2007;318(5849):420-426.

85. Lee S, Kasif S, Weng Z, Cantor CR. Quantitative analysis of single nucleotide polymorphisms within copy number variation. *PloS One*. 2008;3(12):e3906.

86. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 2009;10:451-481.

87. Felekkis K, Voskarides K, Dweep H, Sticht C, Gretz N, Deltas C. Increased number of microRNA target sites in genes encoded in CNV regions. evidence for an evolutionary genomic interaction. *Mol Biol Evol*. 2011;28(9):2421-2424.

88. Lee C, Scherer SW. The clinical context of copy number variation in the human genome. *Expert Rev Mol Med*. 2010;12:e8.

89. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet*. 2005;38(1):75-81.

90. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. *Nat Rev Genet*. 2006;7(2):85-97.

91. Lupski JR, Stankiewicz P. Genomic disorders: Molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS Genetics*. 2005;1(6):e49.

92. Stranger BE, Forrest MS, Dunning M, et al. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*. 2007;315(5813):848-853.

93. Crutchley JL, Wang XQD, Ferraiuolo MA, Dostie J. Chromatin conformation signatures: Ideal human disease biomarkers? *Biomark Med*. 2010;4(4):611-629.

94. Li X, Heyer W. Homologous recombination in DNA repair and DNA damage tolerance. *Cell Res*. 2008;18(1):99-113.

95. Hastings P, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet*. 2009;10(8):551-564.

96. Llorente B, Smith CE, Symington LS. Break-induced replication: What is it and what is it for? *Cell Cycle*. 2008;7(7):859.

97. Arnaudeau C, Lundin C, Helleday T. DNA double-strand breaks associated with replication forks are predominantly repaired by homologous recombination involving an exchange mechanism in mammalian cells. *J Mol Biol*. 2001;307(5):1235-1245.

98. McVey M, Lee SE. MMEJ repair of double-strand breaks (director's cut): Deleted sequences and alternative endings. *Trends Genet*. 2008;24(11):529-538.

99. Liskay RM, Letsou A, Stachelek JL. Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics*. 1987;115(1):161-167.

100. Sugawara N, Ira G, Haber JE. DNA length dependence of the single-strand annealing pathway and the role of saccharomyces cerevisiae RAD59 in double-strand break repair. *Mol Cell Biol*. 2000;20(14):5300-5309.

101. Roth DB, Wilson JH. Nonhomologous recombination in mammalian cells: Role for short sequence homologies in the joining reaction. *Mol Cell Biol*. 1986;6(12):4295-4304.

102. Derbyshire MK, Epstein LH, Young CS, Munz PL, Fishel R. Nonhomologous recombination in human cells. *Mol Cell Biol*. 1994;14(1):156-169.

103. Murnane JP. Telomere dysfunction and chromosome instability. *Mutat Res*. 2012;730(1):28-36.

104. Richards RI, Sutherland GR. Dynamic mutations: A new class of mutations causing human disease. *Cell*. 1992;70(5):709-712.

105. Lovett ST, Feschenko VV. Stabilization of diverged tandem repeats by mismatch repair: Evidence for deletion formation via a misaligned replication intermediate. *Proc Natl Acad Sci U S A*. 1996;93(14):7120-7124.

106. Lee JA, Carvalho C, Lupski JR. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*. 2007;131(7):1235-1247.

107. Gu W, Zhang F, Lupski JR. Mechanisms for human genomic rearrangements. *Pathogenetics*. 2008;1(1):4.

108. Hyrien O. Mechanisms and consequences of replication fork arrest. *Biochimie*. 2000;82(1):5-17.

109. Branzei D, Foiani M. The DNA damage response during DNA replication. *Curr Opin Cell Biol*. 2005;17(6):568-575.

110. Sharp AJ, Locke DP, McGrath SD, et al. Segmental duplications and copy-number variation in the human genome. *AJHG*. 2005;77(1):78-88.

111. Zhang F, Khajavi M, Connolly AM, Towne CF, Batish SD, Lupski JR. The DNA replication FoSTeS/MMBIR mechanism can generate genomic, genic and exonic complex rearrangements in humans. *Nat Genet*. 2009;41(7):849-853.

112. Cheung J, Estivill X, Khaja R, et al. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*. 2003;4(4):R25.

113. Duan J, Zhang J, Deng H, Wang Y. Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PloS One*. 2013;8(3):e59128.

114. Speicher MR, Carter NP. The new cytogenetics: Blurring the boundaries with molecular biology. *Nat Rev Genet*. 2005;6(10):782-792.

115. Cowell JK, Matsui S, Wang YD, et al. Application of bacterial artificial chromosome array-based comparative genomic hybridization and spectral karyotyping to the analysis of glioblastoma multiforme. *Cancer Genet Cytogenet*. 2004;151(1):36-51.

116. Strachan, Tom & Read, Andrew. Human molecular genetics. In: 3rd ed. New York, USA: Garland Science; 2004:374-564.

117. Cheung V, Nowak N, Jang W, et al. Integration of cytogenetic landmarks into the draft sequence of the human genome. *Nature*. 2001;409(6822):953-958.

118. O'Connor M, Peifer M, Bender W. Construction of large DNA segments in escherichia coli. *Science*. 1989;244(4910):1307-1312.

119. Shizuya H, Kouros-Mehr H. The development and applications of the bacterial artificial chromosome cloning system. *Keio J Med*. 2001;50(1):26-30.

120. Pinkel D, Albertson DG. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*. 2005;37:S11-S17.

121. Massion PP, Kuo WL, Stokoe D, et al. Genomic copy number analysis of non-small cell lung cancer using array comparative genomic hybridization: Implications of the phosphatidylinositol 3-kinase pathway. *Cancer Res*. 2002;62(13):3636-3640.

122. Ou Z, Kang SH, Shaw CA, et al. Bacterial artificial chromosome-emulation oligonucleotide arrays for targeted clinical array-comparative genomic hybridization analyses. *Genet Med*. 2008;10(4):278-289.

123. Tchinda J, Lee C. Detecting copy number variation in the human genome using comparative genomic hybridization. *BioTechniques*. 2006;41(4):385-387.

124. Shinawi M, Cheung SW. The array CGH and its clinical applications. *Drug Discov Today*. 2008;13(17):760-770.

125. Marquis-Nicholson R, Aftimos S, Hayes I, George A, Love DR. Array comparative genomic hybridisation: A new tool in the diagnostic genetic armoury. *NZMJ*. 2010;123(1318).

126. Weiss MM, Hermsen MA, Meijer GA, et al. Comparative genomic hybridisation. *Mol Pathol*. 1999;52(5):243-251.

127. Agilent Technologies Canada Inc. Human genome CNV microarrays. http://www.genomics.agilent.com/en/product.jsp?cid=AG-PT-110&tabId=AG-PR-1081&_requestid=24158. Updated 2015. Accessed 05/04, 2015.

128. LaFramboise T. Single nucleotide polymorphism arrays: A decade of biological, computational and technological advances. *Nucleic Acids Res*. 2009;37(13):4181-4193.

129. Steemers FJ, Chang W, Lee G, Barker DL, Shen R, Gunderson KL. Whole-genome genotyping with the single-base extension assay. *Nat Methods*. 2006;3(1):31-33.

130. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. *Brief Funct Genomic Proteomic*. 2009;8(5):353-366.

131. O'Keefe C, McDevitt MA, Maciejewski JP. Copy neutral loss of heterozygosity: A novel chromosomal lesion in myeloid malignancies. *Blood*. 2010;115(14):2731-2739.

132. Carter NP. Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat Genet*. 2007;39:S16-S21.

133. Haraksingh RR, Abyzov A, Gerstein M, Urban AE, Snyder M. Genome-wide mapping of copy number variation in humans: Comparative analysis of high resolution array platforms. *PLoS One*. 2011;6(11):e27859.

134. Vermeesch JR, Brady PD, Sanlaville D, Kok K, Hastings RJ. Genome-wide arrays: Quality criteria and platforms to be used in routine diagnostics. *Hum Mutat*. 2012;33(6):906-915.

135. Illumina. Genome-wide DNA analysis BeadChips, offering a combination of powerful content and unprecedented flexibility for experimental design. http://support.illumina.com/content/dam/illumina-marketing/documents/products/brochures/datasheet_omni_whole-genome_arrays.pdf. Updated 2010, 2011. Accessed 09/11, 2014.

136. Bae JS, Cheong HS, Kim LH, et al. Identification of copy number variations and common deletion polymorphisms in cattle. *BMC Genomics*. 2010;11:232.

137. Illumina. Genome-wide DNA analysis BeadChips. http://www.dkfz.de/gpcf/fileadmin/downloads/Genotyping/datasheet_infiniumhd.pdf. Updated 2010. Accessed 12/03, 2013.

138. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52-58.

139. Illumina. HumanOmni quad BeadChip.

http://WWW.illumina.com/documents/products/datasheets/datasheet_humanomni1_quad.pdf. Updated 2009. Accessed 09/11, 2014.

140. UCSC. Schema for SNP arrays - SNP genotyping arrays and SNP arrays (snpArray) track description. https://cgwb.nci.nih.gov/cgi-bin/hgTables?db=hg19&hgta_group=varRep&hgta_track=snpArray&hgta_table=snpArrayIllumina1M&hgta_doSchema=describe+table+schema. Updated 2014. Accessed 09/11, 2014.

141. Gunderson KL, Steemers FJ, Ren H, et al. Whole-genome genotyping. *Meth Enzymol*. 2006;410:359-376.

142. Wineinger NE, Kennedy RE, Erickson SW, Wojczynski MK, Bruder C, Tiwari H. Statistical issues in the analysis of DNA copy number variations. *Int J Comput Biol Drug Des*. 2008;1(4):368-395.

143. Smyth GK, Speed T. Normalization of cDNA microarray data. *Methods*. 2003;31(4):265-273.

144. Yau C, Holmes CC. CNV discovery using SNP genotyping arrays. *Cytogenet Genome Res*. 2008;123(1-4):307-312.

145. Illumina. Illumina genotyping data normailzation methods.

http://dnatech.genomecenter.ucdavis.edu/wpcontent/uploads/2013/06/illumina_gt_normalization.pdf. Updated 2006. Accessed 09/26, 2014.

146. Teo YY, Inouye M, Small KS, et al. A genotype calling algorithm for the illumina BeadArray platform. *Bioinformatics*. 2007;23(20):2741-2746.

147. Vermeesch JR, Fiegler H, de Leeuw N, et al. Guidelines for molecular karyotyping in constitutional genetic diagnosis. *EJHG*. 2007;15(11):1105-1114.

148. Viterbi AJ. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory*. 1967;13(2):260-269.

149. Baum LE, Petrie T, Soules G, Weiss N. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The Annuals of Mathematical Statistics*. 1970:164-171.

150. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. *Nucleic Acids Res*. 2010;38(9):e105.

151. Zhang X, Du R, Li S, Zhang F, Jin L, Wang H. Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics*. 2014;15(1):1.

152. Marenne G, Rodríguez-Santiago B, Closas MG, et al. Assessment of copy number variation using the illumina infinium 1M SNP-array: A comparison of methodological

approaches in the spanish bladder cancer/EPICURO study. *Hum Mutat*. 2011;32(2):240-248.

153. Lin P, Hartz SM, Wang JC, et al. Copy number variation accuracy in genome-wide association studies. *Hum Hered*. 2011;71(3):141-147.

154. Kim SY, Kim JH, Chung YJ. Effect of combining multiple CNV defining algorithms on the reliability of CNV calls from SNP genotyping data. *Genomics Inform*. 2012;10(3):194-199.

155. Ghani M, Pinto D, Lee JH, et al. Genome-wide survey of large rare copy number variants in alzheimer's disease among caribbean hispanics. *G3 (Bethesda)*. 2012;2(1):71-78.

156. Griswold AJ, Ma D, Cukier HN, et al. Evaluation of copy number variations reveals novel candidate genes in autism spectrum disorder-associated pathways. *Hum Mol Genet*. 2012;21(15):3513-3523.

157. Ukkola-Vuoti L, Kanduri C, Oikkonen J, et al. Genome-wide copy number variation analysis in extended families and unrelated individuals characterized for musical aptitude and creativity in music. *PloS One*. 2013;8(2):e56356.

158. Teo S, Ku C, Naidoo N, et al. A population-based study of copy number variants and regions of homozygosity in healthy swedish individuals. *J Hum Genet*. 2011;56(7):524-533.

159. Ionita-Laza I, Rogers AJ, Lange C, Raby BA, Lee C. Genetic association analysis of copy-number variation (CNV) in human disease pathogenesis. *Genomics*. 2009;93(1):22-26.

160. Jin G, Sun J, Liu W, et al. Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer. *Carcinogenesis*. 2011;32(7):1057-1062.

161. Daigo Y, Nishiwaki T, Kawasoe T, Tamari M, Tsuchiya E, Nakamura Y. Molecular cloning of a candidate tumor suppressor gene, DLC1, from chromosome 3p21.3. *Cancer Res*. 1999;59(8):1966-1972.

162. Gonzalez E, Kulkarni H, Bolivar H, et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*. 2005;307(5714):1434-1440.

163. Kudo S, Kashida H, Nakajima T, Tamura S, Nakajo K. Endoscopic diagnosis and treatment of early colorectal cancer. *World J Surg*. 1997;21(7):694-701.

164. The American Cancer Society. Colorectal cancer. http://www.cancer.org/acs/groups/cid/documents/webcontent/003096-pdf.pdf. Updated 2013. Accessed 06/12, 2013.

165. Colorectal Cancer of Canada. What are the symptoms of colorectal cancer? http://www.colorectal-cancer.ca/en/just-the-facts/symptoms/. Updated 2013. Accessed 06/20, 2013.

166. Warmkessel JH. Caring for a patient with colon cancer. *Nursing2013*. 1997;27(4):34-40.

167. Gore RM. Gastrointestinal. In: Ronald L. Eisenberg, ed. *The right imaging study.* 3rd ed. Springer; 2008:173-272.

168. Ferlay J, Shin H, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. 2010;127(12):2893-2917.

169. Center MM, Jemal A, Smith RA, Ward E. Worldwide variations in colorectal cancer. *CA Cancer J Clin*. 2009;59(6):366-378.

170. Huxley RR, Ansary-Moghaddam A, Clifton P, Czernichow S, Parr CL, Woodward M. The impact of dietary and lifestyle risk factors on risk of colorectal cancer: A quantitative overview of the epidemiological evidence. *Int J cancer*. 2009;125(1):171-180.

171. Porta MS, Greenland S, Hernán M, Silva IDS, Last JM. *A dictionary of epidemiology.* 6th ed. Oxford University Press; 2014.

172. Mayo Clinic. Colon polyps. http://www.mayoclinic.com/health/colon-polyps/DS00511/DSECTION=causes. Updated 2014. Accessed 09/26, 2014.

173. Johns Hopkins Medicine. From polyp to cancer, colorectal cancer. http://www.hopkinscoloncancercenter.org/CMS/CMS_Page.aspx?CurrentUDV=59&CMS_Page_ID=0B34E9BE-5DE6-4CB4-B387-4158CC924084. Updated 2014. Accessed 09/20, 2014.

174. Hyman NH, Anderson P, Blasyk H. Hyperplastic polyposis and the risk of colorectal cancer. *Dis Colon Rectum*. 2004;47(12):2101-2104.

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

176. Potter JD, Slattery ML, Bostick RM, Gapstur SM. Colon cancer: A review of the epidemiology. *Epidemiol Rev*. 1993;15(2):499-545.

177. Campos F, Logullo Waitzberg A, Kiss D, Waitzberg D, Habr-Gama A, Gama-Rodrigues J. Diet and colorectal cancer: Current evidence for etiology and prevention. *Nutr Hosp*. 2005;20(1):18-25.

178. Grahn SW, Varma MG. Factors that increase risk of colon polyps. *Clin Colon Rectal Surg*. 2008;21(4):247-255.

179. Thune I, Lund E. Physical activity and risk of colorectal cancer in men and women. *Br J Cancer*. 1996;73(9):1134-1140.

180. Slattery ML. Physical activity and colorectal cancer. *Sports Medicine*. 2004;34(4):239-252.

181. DeCosse J, Ngoi S, Jacobson J, Cennerazzo W. Gender and colorectal cancer. *ECP*. 1993;2(2):105-116.

182. Center MM, Jemal A, Ward E. International trends in colorectal cancer incidence rates. *Cancer Epidemiol Biomarkers Prev*. 2009;18(6):1688-1694.

183. Boyle P, Langman JS. ABC of colorectal cancer: Epidemiology. *BMJ*. 2000;321(7264):805-808.

184. Church J, McGannon E. Family history of colorectal cancer. *Dis Colon Rectum*. 2000;43(11):1540-1544.

185. De la Chapelle A. Genetic predisposition to colorectal cancer. *Nat Rev Cancer*. 2004;4(10):769-780.

186. Canadian Cancer Society. Risk factors for colorectal cancer. http://www.cancer.ca/en/cancer-information/cancer-type/colorectal/risks/?region=on#Family_history#ixzz3LycTh2Q8. Updated 2014. Accessed 12/15, 2014.

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

188. Naccarati A, Pardini B, Hemminki K, Vodicka P. Sporadic colorectal cancer and individual susceptibility: A review of the association studies investigating the role of DNA repair genetic polymorphisms. *Mutation Research*. 2007;635(2):118-145.

189. Fearon ER. Molecular genetics of colorectal cancer. *Annu Rev Pathol Mech Dis.* 2011;6:479-507.

190. Robertson JM. Early diagnosis and treatment of cancer: Colorectal cancer. *Int J Radiat Oncol Biol Phys*. 2011;79(5):1597-1597.

191. Kinzler KW, Vogelstein B. Lessons from hereditary colorectal cancer. *Cell*. 1996;87(2):159-170.

192. Hamilton SR, Liu B, Parsons RE, et al. The molecular basis of turcot's syndrome. *N Engl J Med*. 1995;332(13):839-847.

193. Lebrun C, Olschwang S, Jeannin S, Vandenbos F, Sobol H, Frenay M. Turcot syndrome confirmed with molecular analysis. *Eur J Neurol*. 2007;14(4):470-472.

194. Shin SH, Yu EJ, Lee Y, Song Y, Seong M, Park SS. Characteristics of hereditary nonpolyposis colorectal cancer patients with double primary cancers in endometrium and colorectum. *Obstet Gynecol Sc*. 2015;58(2):112-116.

195. Dowty JG, Win AK, Buchanan DD, et al. Cancer risks for MLH1 and MSH2 mutation carriers. *Hum Mutat*. 2013;34(3):490-497.

196. Lynch H, Lynch P, Lanspa S, Snyder C, Lynch J, Boland C. Review of the lynch syndrome: History, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clin Genet*. 2009;76(1):1-18.

197. Hare HH, Mahendraker N, Sarwate S, Tangella K. Muir-torre syndrome: A rare but important disorder. *Cutis*. 2008;82(4):252-256.

198. Bhaijee F, Brown AS. Muir-torre syndrome. *Arch Pathol Lab Med*. 2014;138(12):1685.

199. Luber AJ, Zeichner JA. Muir-torre syndrome. In: *Acneiform eruptions in dermatology.* Springer; 2014:215-219.

200. Nieminen TT, Abdel–Rahman WM, Ristimäki A, et al. BMPR1A mutations in hereditary nonpolyposis colorectal cancer without mismatch repair deficiency. *Gastroenterology*. 2011;141(1):e23-e26.

201. Nieminen TT, O'Donohue M, Wu Y, et al. Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. *Gastroenterology*. 2014;147(3):595-598. e5.

202. Howe JR, Roth S, Ringold JC, et al. Mutations in the SMAD4/DPC4 gene in juvenile polyposis. *Science*. 1998;280(5366):1086-1088.

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

204. Cruz-Correa M. Very high risk of cancer in familial peutz-jeghers syndrome. *Gastroenterology*. 2000;119(6):1447-1453.

205. Kopacova M, Tacheci I, Rejchrt S, Bures J. Peutz-jeghers syndrome: Diagnostic and therapeutic approach. *World J Gastroenterol*. 2009;15(43):5397-5408.

206. Beggs AD, Latchford AR, Vasen HF, et al. Peutz-jeghers syndrome: A systematic review and recommendations for management. *Gut*. 2010;59(7):975-986.

207. Nielsen M, Morreau H, Vasen HF, Hes FJ. MUTYH-associated polyposis (MAP). *Crit Rev Oncol*. 2011;79(1):1-16.

208. Lubbe SJ, Di Bernardo MC, Chandler IP, Houlston RS. Clinical implications of the colorectal cancer risk associated with MUTYH mutation. *J Clin Oncol*. 2009;27(24):3975-3980.

209. Rubio C, Stemme S, Jaramillo E, Lindblom A. Hyperplastic polyposis coli syndrome and colorectal carcinoma. *Endoscopy*. 2006;38(03):266-270.

210. Farooq A, Walker L, Bowling J, Audisio R. Cowden syndrome. *Cancer Treat Rev*. 2010;36(8):577-583.

211. Chung TP, Fleshman JW. The genetics of sporadic colon cancer. *Semin Colon Rectal Surg*. 2004;15(3):128-135.

212. Dai J, Gu J, Huang M, et al. GWAS-identified colorectal cancer susceptibility loci associated with clinical outcomes. *Carcinogenesis*. 2012;33(7):1327-1331.

213. Houlston RS, Cheadle J, Dobbins SE, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26. 2, 12q13. 13 and 20q13. 33. *Nat Genet*. 2010;42(11):973-977.

214. Dunlop MG, Dobbins SE, Farrington SM, et al. Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nat Genet*. 2012;44(7):770-776.

215. Hutter CM, Chang-Claude J, Slattery ML, et al. Characterization of gene-environment interactions for colorectal cancer susceptibility loci. *Cancer Res*. 2012;72(8):2036-2044.

216. Horpaopan S, Spier I, Zink AM, et al. Genome-wide CNV analysis in 221 unrelated patients and targeted high-throughput sequencing reveal novel causative candidate genes for colorectal adenomatous polyposis. *Int J Cancer*. 2015;136(6):E578-E589.

217. Park RW, Kim TM, Kasif S, Park PJ. Identification of rare germline copy number variations over-represented in five human cancer types. *Mol Cancer*. 2015;14:25-015-0292-6.

218. Venkatachalam R, Verwiel ET, Kamping EJ, et al. Identification of candidate predisposing copy number variants in familial and early-onset colorectal cancer patients. *Int J Cancer*. 2011;129(7):1635-1642.

219. Chen W, Yuan L, Cai Y, et al. Identification of chromosomal copy number variations and novel candidate loci in hereditary nonpolyposis colorectal cancer with mismatch repair proficiency. *Genomics*. 2013;102(1):27-34.

220. Thean L, Loi C, Ho K, Koh P, Eu K, Cheah P. Genome-wide scan identifies a copy number variable region at 3q26 that regulates PPM1L in APC mutation-negative familial colorectal cancer patients. *Genes, Chromosomes Cancer*. 2010;49(2):99-106.

221. Seshagiri S, Stawiski EW, Durinck S, et al. Recurrent R-spondin fusions in colon cancer. *Nature*. 2012;488(7413):660-664.

222. Pineau P, Nagai H, Prigent S, et al. Identification of three distinct regions of allelic deletions on the short arm of chromosome 8 in hepatocellular carcinoma. *Oncogene*. 1999;18(20):3127-3134.

223. Pils D, Horak P, Gleiss A, et al. Five genes from chromosomal band 8p22 are significantly down-regulated in ovarian carcinoma. *Cancer*. 2005;104(11):2417-2429.

224. Chaib H, MacDonald JW, Vessella RL, et al. Haploinsufficiency and reduced expression of genes localized to the 8p chromosomal region in human prostate tumors. *Genes Chromosomes Cancer*. 2003;37(3):306-313.

225. Farrington SM, Cunningham C, Boyle SM, Wyllie AH, Dunlop MG. Detailed physical and deletion mapping of 8p with isolation of YAC clones from tumour suppressor loci involved in colorectal cancer. *Oncogene*. 1996;12(8):1803-1808.

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

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

228. Uddin M, Sturge M, Rahman P, Woods MO. Autosome-wide copy number variation association analysis for rheumatoid arthritis using the WTCCC high-density SNP genotype data. *J Rheumatol*. 2011;38(5):797-801.

229. Zheng X, Shaffer JR, McHugh CP, et al. Using family data as a verification standard to evaluate copy number variation calling strategies for genetic association studies. *Genet Epidemiol*. 2012;36(3):253-262.

230. Colella Stefano. QuantiSNP algorithm download. https://sites.google.com/site/quantisnp/downloads. Updated 2011. Accessed 04/19, 2013.

231. Diskin SJ, Li M, Hou C, et al. Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res*. 2008;36(19):e126.

232. Kuningas M, Estrada K, Hsu YH, et al. Large common deletions associate with mortality at old age. *Hum Mol Genet*. 2011;20(21):4290-4296.

233. Wang Kai. PennCNV: Copy number variations detection download. http://www.openbioinformatics.org/penncnv/. Updated 2011. Accessed 02/10, 2013.

234. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210.

235. Illumina support. HumanOmni1-Quad_v1-0_B-H_MappingInformation.txt. http://support.illumina.com/search.ilmn?search=HumanOmni1-Quad%20v1-0%20B-H%20Mapping%20Information&Pg=1&ilmn_search_btn.x=1&filter=. Updated 2014. Accessed 02/05, 2014.

236. Buizer-Voskamp JE, Muntjewerff J, Strengman E, et al. Genome-wide analysis shows increased frequency of copy number variation deletions in dutch schizophrenia patients. *Biol Psychiatry*. 2011;70(7):655-662.

237. Jiang L, Jiang J, Yang J, et al. Genome-wide detection of copy number variations using high-density SNP genotyping platforms in holsteins. *BMC Genomics*. 2013;14:131.

238. UCSC. gc5Base.txt file download. http://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/gc5Base.txt.gz. Updated 2009. Accessed 01/28, 2014.

239. Karolchik D, Barber GP, Casper J, et al. The UCSC genome browser database: 2014 update. *Nucleic Acids Res*. 2014;42(Database issue):D764-70.

240. Pinto D, Pagnamenta AT, Klei L, et al. Functional impact of global rare copy number variation in autism spectrum disorders. *Nature*. 2010;466(7304):368-372.

241. Nag A, Bochukova EG, Kremeyer B, et al. CNV analysis in tourette syndrome implicates large genomic rearrangements in COL8A1 and NRXN1. *PloS One*. 2013;8(3):e59061.

242. Degenhardt F, Priebe L, Herms S, et al. Association between copy number variants in 16p11. 2 and major depressive disorder in a german case–control sample. *Am J Med Genet*. 2012;159(3):263-273.

243. Priebe L, Degenhardt F, Strohmaier J, et al. Copy number variants in german patients with schizophrenia. *PloS One*. 2013;8(7):e64035.

244. Marenne G, Real FX, Rothman N, et al. Genome-wide CNV analysis replicates the association between GSTM1 deletion and bladder cancer: A support for using continuous measurement from SNP-array data. *BMC Genomics*. 2012;13:326.

245. Need AC, Ge D, Weale ME, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genetics*. 2009;5(2):e1000373.

246. Wineinger NE, Pajewski NM, Kennedy RE, et al. Characterization of autosomal copy-number variation in african americans: The HyperGEN study. *Eur J Hum Genet*. 2011;19(12):1271-1275.

247. Talseth-Palmer BA, Holliday EG, Evans TJ, et al. Continuing difficulties in interpreting CNV data: Lessons from a genome-wide CNV association study of australian HNPCC/lynch syndrome patients. *BMC Med Genomics*. 2013;6:10.

248. Fernandez TV, Sanders SJ, Yurkiewicz IR, et al. Rare copy number variants in tourette syndrome disrupt genes in histaminergic pathways and overlap with autism. *Biol Psychiatry*. 2012;71(5):392-402.

249. Tropeano M, Ahn JW, Dobson RJ, et al. Male-biased autosomal effect of 16p13. 11 copy number variation in neurodevelopmental disorders. *PLoS One*. 2013;8(4):e61365.

250. Safran M, Chalifa-Caspi V, Shmueli O, et al. Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res*. 2003;31(1):142-146.

251. Bae JS, Cheong HS, Park BL, et al. Genome-wide profiling of structural genomic variations in korean HapMap individuals. *PloS One*. 2010;5(7):e11417.

252. Campbell CD, Sampas N, Tsalenko A, et al. Population-genetic properties of differentiated human copy-number polymorphisms. *Am J Med Genet*. 2011;88(3):317-332.

253. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. *Nature*. 2010;464(7289):704-712.

254. Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res*. 2014;42(Database issue):D749-55.

255. Mi H, Muruganujan A, Thomas PD. PANTHER in 2013: Modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Res*. 2013;41(Database issue):D377-86.

256. Gene Ontology Consortium. The gene ontology (GO) project in 2006. *Nucleic Acids Res*. 2006;34(Database issue):D322-6.

257. Weizmann Institute of Science. Genecard encyclopedia. www.genecards.org. Updated 10 May 2015. Accessed 06/04, 2015.

258. Marikawa Y, Fujita TC, Alarcón VB. An enhancer-trap LacZ transgene reveals a distinct expression pattern of kinesin family 26B in mouse embryos. *Dev Genes Evol*. 2004;214(2):64-71.

259. Krepischi A, Achatz M, Santos E, et al. Germline DNA copy number variation in familial and early-onset breast cancer. *Breast Cancer Res*. 2012;14(1):R24.

260. Wang Q, Zhao Z, Wang G, et al. High expression of KIF26B in breast cancer associates with poor prognosis. *PloS One*. 2013;8(4).

261. Liu P, Choi YK, Qi RZ. NME7 is a functional component of the gamma-tubulin ring complex. *Mol Biol Cell*. 2014;25(13):2017-2025.

262. Diskin SJ, Bosse K, Mayes PA, et al. Identification of NME7 as a predisposition locus and candidate oncogene in neuroblastoma. *Cancer Res*. 2010;70(8 Supplement):3866-3866.

263. Zabala W, Cruz R, Barreiro-de Acosta M, et al. New genetic associations in thiopurine-related bone marrow toxicity among inflammatory bowel disease patients. *Pharmacogenomics*. 2013;14(6):631-640.

264. Schmit SL, Schumacher FR, Edlund CK, et al. A novel colorectal cancer risk locus at 4q32.2 identified from an international genome-wide association study. *Carcinogenesis*. 2014;35(11):2512-2519.

265. Massagué J, Blain SW, Lo RS. TGFβ signaling in growth control, cancer, and heritable disorders. *Cell*. 2000;103(2):295-309.

266. Xu Y, Pasche B. TGF-beta signaling alterations and susceptibility to colorectal cancer. *Hum Mol Genet*. 2007;16 Spec No 1:R14-20.

267. Paige AJ, Taylor KJ, Taylor C, et al. WWOX: A candidate tumor suppressor gene involved in multiple tumor types. *Proc Natl Acad Sci U S A*. 2001;98(20):11417-11422.

268. Sun PC, Uppaluri R, Schmidt AP, et al. Transcript map of the 8p23 putative tumor suppressor region. *Genomics*. 2001;75(1):17-25.

269. Ali Hassan N, Mokhtar N, Kok Sin T, et al. Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. *PloS One*. 2014;9(4):e92553.

270. Giles RH, van Es JH, Clevers H. Caught up in a wnt storm: Wnt signaling in cancer. *BBA*. 2003;1653(1):1-24.

271. Bienz M, Clevers H. Linking colorectal cancer to wnt signaling. *Cell*. 2000;103(2):311-320.

272. Grothey A, Galanis E. Targeting angiogenesis: Progress with anti-VEGF treatment with large molecules. *Nat Rev Clin Oncol*. 2009;6(9):507-518.

273. Sakurai T, Kudo M. Signaling pathways governing tumor angiogenesis. *Oncology*. 2011;81 Suppl 1:24-29.

274. Ling H, Vincent K, Pichler M, et al. Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*. 2015.

275. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS. Non-coding RNAs: Regulators of disease. *J Pathol*. 2010;220(2):126-139.

276. Masson AL, Talseth-Palmer BA, Evans T, et al. Copy number variation in hereditary non-polyposis colorectal cancer. *Genes*. 2013;4(4):536-555.

APPENDICES

Appendix A: Copyright approval to adapt and use the figure from *Lee et al (2010)*.

RightsLink

Copyright Clearance Center

Permissions Request Requires Your Acceptance

Dear Mr. Salem Werdyani,

Cambridge University Press has approved your recent permission request - see the details below. Prior to reusing this content, you must accept the license fee and terms.

To accept or decline this order, please click the link below to open RightsLink.
https://s100.copyright.com/CustomerAdmin/FC.jsp?ref=56f3e530-9138-4093-9b3b-cb1a32f0d11b&pName=CUP

(If the link above is displaying on two lines, it may not open your browser window properly. Copy and paste the entire link into your browser address field and try again.)

Order Details

Licensee:	Salem Werdyani
Order Date:	Jul 30, 2015
Order Number:	501033457
Publication:	Expert Reviews in Molecular Medicine
Title:	The clinical context of copy number variation in the human genome
Type of Use:	Dissertation/Thesis

Latest Comment: Your RightsLink request has been approved.

This service provides permission for reuse only. If you require a copy of the article you are using, please select the "Pay per View" option (via the...

Note: Payee for this order is Copyright Clearance Center.

B.3:v5.7

+1-855-239-3415 / Tel: +1-978-646-2777
customercare@copyright.com
http://www.copyright.com

Copyright Clearance Center RightsLink®

Account Info | Help | Live Chat

CAMBRIDGE JOURNALS

Title: The clinical context of copy number variation in the human genome

Author: Charles Lee and Stephen W. Scherer

Publication: Expert Reviews in Molecular Medicine

Publisher: Cambridge University Press

Date: Mar 9, 2010

Logged in as:
Salem Werdyani
Account #:
3000945861

LOGOUT

Order Completed

Thank you for your order.

This Agreement between Salem Werdyani ("You") and Cambridge University Press ("Cambridge University Press") consists of your order details and the terms and conditions provided by Cambridge University Press and Copyright Clearance Center.

License number	Reference confirmation email for license number
License date	Jul 30, 2015
Licensed content publisher	Cambridge University Press
Licensed content publication	Expert Reviews in Molecular Medicine
Licensed content title	The clinical context of copy number variation in the human genome
Licensed content author	Charles Lee and Stephen W. Scherer
Licensed content date	Mar 9, 2010
Volume number	12
Issue number	-1
Type of Use	Dissertation/Thesis
Requestor type	Not-for-profit
Portion	Full article
Order reference number	None
Territory for reuse	North America Only
Title of your thesis / dissertation	Characteristics of the germ-line Copy Number Variations (CNVs) in colorectal cancer patients
Expected completion date	Jan 2016
Estimated size(pages)	125
Billing Type	Invoice
Billing address	Salem Werdyani 300 PRINCE PHILIP DRIVE St.John's, NL A1B 3V6 Canada Attn: Salem Werdyani
Tax (0.00%)	0.00 USD
Total	0.00 USD

Appendix B: Copyright approval to adapt and use the figure from *Hastings et al (2009)*.

NATURE PUBLISHING GROUP LICENSE
TERMS AND CONDITIONS

Oct 05, 2015

This is a License Agreement between Salem Werdyani ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3722101191587
License date	Oct 04, 2015
Licensed content publisher	Nature Publishing Group
Licensed content publication	Nature Reviews Genetics
Licensed content title	Mechanisms of change in gene copy number
Licensed content author	P. J. Hastings, James R. Lupski, Susan M. Rosenberg and Grzegorz Ira
Licensed content date	Aug 1, 2009
Volume number	10
Issue number	8
Type of Use	reuse in a dissertation / thesis
Requestor type	academic/educational
Format	electronic
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	4
High-res required	no
Figures	Mechanisms of homologous recombination, change in copy number by homologous recombination, the breakage-fusion-bridge cycle, and Replicative mechanisms for non-homologous structural change.
Author of this NPG article	no
Your reference number	None
Title of your thesis / dissertation	Characteristics of the germ-line Copy Number Variations (CNVs) in colorectal cancer patients
Expected completion date	Jan 2016
Estimated size (number of pages)	125
Total	0.00 USD
Terms and Conditions	

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use,subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version.Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run).NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)
   For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

   Note: For republication from the *British Journal of Cancer*, the following credit lines apply.
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)For AOP papers, the credit line should read:
   Reprinted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

   Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

   Note: For adaptation from the *British Journal of Cancer*, the following credit line applies.
   Adapted by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information.Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

Note: For translation from the *British Journal of Cancer*, the following credit line applies.
Translated by permission from Macmillan Publishers Ltd on behalf of Cancer Research UK: [JOURNAL NAME] (reference citation), copyright (year of publication)

We are certain that all parties will benefit from this agreement and wish you the best in the use of this material. Thank you.
Special Terms:
v1.1

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix C: Copyright approval to adapt and use the figure from Illumina.

Image Use Agreement

By accessing the [Illumina Image Store] (the "**Site**") and downloading, or making any use of any image ("**Image**") on the Site, you agree to abide by and be bound by the terms of this Image Use Agreement (this "**Agreement**").

<u>Ownership</u>
The Site is owned by Illumina, Inc. ("**Illumina**"). Unless otherwise indicated, all of the content featured or displayed on the Site, including all Images, is owned by Illumina, its licensors, or its third-party partners. All content on the Site is protected by copyright, trade dress, moral rights, trademark, and other laws relating to the protection of intellectual property.

<u>Grant of License and Use of the Site and Images</u>
Subject to the terms of this Agreement, Illumina grants you a limited, revocable, non-exclusive license to access, download, copy, and/or use the Images on the Site subject to the terms of this Agreement. The right and license granted herein is personal to you and is non-sublicensable, non-transferable and non-assignable by you to any other person or entity.

When using an Image you <u>must</u>: (a) always include a credit line as follows: "Courtesy of Illumina, Inc." and (b) notify Illumina of any misuse or unauthorized use of the Image of which you become aware.

When using an Image you must <u>not</u>: (a) use any Image for any commercial purpose, including the promotion of any product or technology, or for purposes of product comparisons; (b) make the Image available in a manner intended to allow or invite a third party to download, extract, or otherwise access the Image as a standalone file; (c) alter or modify the Image, in whole or in part, for any reason (except for re-sizing as appropriate); (d) delete, alter, or obfuscate any proprietary legends or attributions relating to the Image, including any trademark or copyright notations; (e) use an Image in a defamatory, offensive, or otherwise unlawful manner; (f) incorporate an Image into a logo, trademark, or service mark, or use an Image in any way to suggest an affiliation with or endorsement of any product, service, or commercial or non-commercial enterprise.

You are further specifically prohibited from: (a) using any data mining, robots, or similar data gathering or extraction methods on the Site; (b) reverse engineering, altering, or modifying any part of the Site or Images; (c) circumventing, disabling, or otherwise interfering with security-related features of the Site or on the Images; (d) selling, licensing, leasing, or in any way commercializing the Images; and (e) using the Site or any Image other than allowed by these terms.

Unauthorized use of this Site or any Image or any breach of this Agreement will cause Illumina irreparable harm and may violate applicable laws including without limitation copyright and trademark laws, the laws of privacy and publicity, and applicable communications regulations and statutes. You represent and warrant that you will comply with all applicable laws and regulations, including, without limitation, those relating to the Internet, data, e-mail, privacy, and the transmission of technical data exported from the United States or the country in which you reside.

No Warranty
The Site and Images are provided by Illumina on an "as is" basis, without warranty of any kind, including non-infringement or ownership. Illumina does not grant any right nor make any warranty with regard to the use of names, people, moral rights, trademarks, trade dress, logos, or designs depicted in any Images, and you shall be solely responsible for determining whether release(s) is/are required for your requested use, and for obtaining any required releases if not already obtained by Illumina.

Indemnity
You hereby agree to indemnify, defend, and hold harmless Illumina from and against any and all claims, actions, proceedings, suits, liabilities, damages, penalties, fines, losses, or expenses, including attorneys' fees and costs, arising out of or in any way connected with your use of the Site or any Image in violation or breach or alleged violation or breach of any term in this Agreement.

Severability
If any provision of this Agreement is deemed unlawful, void, or for any reason unenforceable, then that provision will be deemed severable from the Agreement and will not affect the validity and enforceability of the remaining provisions. The affected provision shall be modified to the extent necessary to make it lawful, not void, or enforceable, as the case may be, in such a manner as comes closest to preserving the intentions of such provision.

Miscellaneous
This is a legal, binding agreement. You represent and warrant that you have the full right, power and authority to enter into this Agreement and be bound by its terms. The terms of this Agreement will be governed and construed in accordance with the laws of the state of California without regard to its conflicts of law provisions. You agree to submit to the jurisdiction of the state and federal courts located in the state of California in connection with any action arising under this Agreement.

Amendments
Illumina may, in its discretion, change these terms at any time and such changes shall be effective immediately and incorporated into this Agreement. It is your responsibility to review any changes to these terms.

Appendix D: Copyright approval to adapt and use the figure from *Teo et al (2007)*.

OXFORD UNIVERSITY PRESS LICENSE
TERMS AND CONDITIONS

Sep 09, 2015

This is a License Agreement between Salem Werdyani ("You") and Oxford University Press ("Oxford University Press") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Oxford University Press, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

License Number	3703880229325
License date	Sep 07, 2015
Licensed content publisher	Oxford University Press
Licensed content publication	Bioinformatics
Licensed content title	A genotype calling algorithm for the Illumina BeadArray platform:
Licensed content author	Yik Y. Teo, Michael Inouye, Kerrin S. Small, Rhian Gwilliam, Panagiotis Deloukas, Dominic P. Kwiatkowski, Taane G. Clark
Licensed content date	10/15/2007
Type of Use	Thesis/Dissertation
Institution name	None
Title of your work	Characteristics of the germ-line Copy Number Variations (CNVs) in colorectal cancer patients
Publisher of your work	n/a
Expected publication date	Jan 2016
Permissions cost	0.00 USD
Value added tax	0.00 USD
Total	0.00 USD
Total	0.00 USD
Terms and Conditions	

STANDARD TERMS AND CONDITIONS FOR REPRODUCTION OF MATERIAL FROM AN OXFORD UNIVERSITY PRESS JOURNAL

1. Use of the material is restricted to the type of use specified in your order details.

2. This permission covers the use of the material in the English language in the following territory: world. If you have requested additional permission to translate this material, the terms and conditions of this reuse will be set out in clause 12.

3. This permission is limited to the particular use authorized in (1) above and does not allow you to sanction its use elsewhere in any other format other than specified above, nor does it

apply to quotations, images, artistic works etc that have been reproduced from other sources which may be part of the material to be used.

4. No alteration, omission or addition is made to the material without our written consent. Permission must be re-cleared with Oxford University Press if/when you decide to reprint.

5. The following credit line appears wherever the material is used: author, title, journal, year, volume, issue number, pagination, by permission of Oxford University Press or the sponsoring society if the journal is a society journal. Where a journal is being published on behalf of a learned society, the details of that society must be included in the credit line.

6. For the reproduction of a full article from an Oxford University Press journal for whatever purpose, the corresponding author of the material concerned should be informed of the proposed use. Contact details for the corresponding authors of all Oxford University Press journal contact can be found alongside either the abstract or full text of the article concerned, accessible from www.oxfordjournals.org Should there be a problem clearing these rights, please contact journals.permissions@oup.com

7. If the credit line or acknowledgement in our publication indicates that any of the figures, images or photos was reproduced, drawn or modified from an earlier source it will be necessary for you to clear this permission with the original publisher as well. If this permission has not been obtained, please note that this material cannot be included in your publication/photocopies.

8. While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by Oxford University Press or by Copyright Clearance Center (CCC)) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and Oxford University Press reserves the right to take any and all action to protect its copyright in the materials.

9. This license is personal to you and may not be sublicensed, assigned or transferred by you to any other person without Oxford University Press's written permission.

10. Oxford University Press reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

11. You hereby indemnify and agree to hold harmless Oxford University Press and CCC, and their respective officers, directors, employs and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant

to this license.

12. Other Terms and Conditions:

v1.4

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.

Appendix E: The list of the highly repetitive DNA sequence regions based on Hg19 (centromere and telomere regions, leukocyte Immunoglobulin-like receptor gene cluster genes, and olfactory receptor genes).

Highly repetitive DNA sequences based on the Hg19	CHR	Start position	End position
leukocyte Immunoglobulin-like receptor gene cluster (234, 235)	2	89156874	89630187
	14	22090057	23021097
	14	105994256	107281230
	22	22385572	23265082
Centromere Regions (235) **Note:** centromere regions were determined by subtracting 100 kbps before and adding 100 kbps after the centromere position of each chromosome.	1	121400000	129000000
	2	90400000	96900000
	3	87800000	94000000
	4	48100000	52800000
	5	46000000	50800000
	6	58600000	63400000
	7	57900000	61800000
	8	43000000	48200000
	9	47200000	50800000
	10	37900000	42400000
	11	51500000	55800000
	12	33200000	38300000
	13	16200000	19600000
	14	16000000	19200000
	15	15700000	20800000
	16	34500000	38700000
	17	22100000	25900000
	18	15300000	19100000
	19	24300000	28700000
	20	25500000	29500000
	21	10800000	14400000
	22	12100000	18000000
Telomere Regions (234, 235) **Note:** each chromosome contains two telomere region. These regions consist of 500 kbps at the start, and 500 kbps at the end of each chromosome. As a result, there are two telomere regions shown in this table for each chromosome.	1	1	500000
	1	248750621	249250621
	2	1	500000
	2	242699373	243199373
	3	1	500000
	3	197522430	198022430
	4	1	500000
	4	190654276	191154276
	5	1	500000

	5	180415260	180915260
	6	1	500000
	6	170615067	171115067
	7	1	500000
	7	158638663	159138663
	8	1	500000
	8	145864022	146364022
	9	1	500000
	9	140713431	141213431
	10	1	500000
	10	135034747	135534747
	11	1	500000
	11	134506516	135006516
	12	1	500000
	12	133351895	133851895
	13	1	500000
	13	114669878	115169878
	14	1	500000
	14	106849540	107349540
	15	1	500000
	15	102031392	102531392
	16	1	500000
	16	89854753	90354753
	17	1	500000
	17	80695210	81195210
	18	1	500000
	18	77577248	78077248
	19	1	500000
	19	58628983	59128983
	20	1	500000
	20	62525520	63025520
	21	1	500000
	21	47629895	48129895
	22	1	500000
	22	50804566	51304566
Olfactory Receptors (OR) Genes (244)	1	158778192	158779125
	1	52452	53395
	1	63015	63884
	1	69091	70005
	1	367659	368594

	1	621099	622034
	1	111396561	111397511
	1	146890732	146891624
	1	146917311	146917866
	1	158368315	158369256
	1	158389721	158390656
	1	158414863	158415804
	1	158435352	158436290
	1	158449701	158450672
	1	158461009	158461947
	1	158484691	158485715
	1	158516921	158517895
	1	158532444	158533394
	1	158548712	158549638
	1	158576229	158577167
	1	158664329	158665299
	1	158669471	158670442
	1	158686961	158687905
	1	158693945	158694826
	1	158712358	158713205
	1	158724711	158725634
	1	158735537	158736472
	1	158746475	158747425
	1	158765787	158766027
	1	159248834	159249791
	1	159283463	159284449
	1	159320841	159321813
	1	159335951	159336876
	1	159375213	159376199
	1	159401994	159402925
	1	159409582	159410508
	1	159504871	159505797
	1	159551413	159552010
	1	159568088	159569014
	1	247614334	247615284
	1	247654430	247655389
	1	247694854	247695813
	1	247751662	247752612
	1	247768888	247769814
	1	247782983	247783861

	1	247830183	247831118
	1	247835420	247836343
	1	247875134	247876057
	1	247886404	247887345
	1	247901917	247902858
	1	247920767	247921708
	1	247938212	247939135
	1	247978105	247979031
	1	247996662	247997590
	1	248004233	248005198
	1	248058889	248059830
	1	248084320	248085255
	1	248097068	248098054
	1	248102191	248102935
	1	248112160	248113095
	1	248128679	248129638
	1	248138044	248138977
	1	248153569	248154493
	1	248166433	248167368
	1	248185250	248186185
	1	248201570	248202505
	1	248223984	248224919
	1	248246938	248247898
	1	248262678	248263613
	1	248285234	248286163
	1	248308450	248309385
	1	248343288	248344328
	1	248366370	248367305
	1	248402231	248403163
	1	248436157	248437116
	1	248457921	248458880
	1	248486935	248487870
	1	248512077	248513012
	1	248524937	248525926
	1	248550910	248551833
	1	248569449	248570402
	1	248604508	248605431
	1	248616099	248617070
	1	248636652	248637605
	1	248651908	248652834

	1	248661611	248661960
	1	248684948	248685895
	1	248712723	248713204
	1	248721848	248722774
	1	248737105	248738058
	1	248756134	248757069
	1	248789482	248790429
	1	248801591	248802559
	1	248813235	248814185
	1	248844673	248845605
	2	71256002	71257026
	2	71264851	71265795
	2	71282268	71283288
	2	96212326	96213321
	2	159710062	159711085
	2	159718857	159719803
	2	159731151	159732179
	2	240968911	240969846
	2	240984497	240985489
	2	241018850	241019777
	2	241048519	241049466
	3	8729920	8730840
	3	75396998	75397914
	3	75405636	75406661
	3	75419564	75420514
	3	75647802	75648824
	3	97771737	97772662
	3	97783316	97784237
	3	97806017	97806943
	3	97823598	97824513
	3	97851542	97852480
	3	97868230	97869159
	3	97887544	97888482
	3	97916055	97916979
	3	97926267	97927182
	3	97940888	97941837
	3	97957196	97958122
	3	97983177	97984103
	3	98001747	98002673
	3	98030757	98031684

	3	98072698	98073660
	3	98109510	98110472
	3	98188421	98189344
	3	98216525	98217472
	3	112243033	112244047
	3	125422219	125423181
	3	125430969	125431913
	3	125443323	125444357
	3	125453079	125454087
	3	125465918	125466952
	3	129740399	129741423
	3	129753332	129754385
	4	3891030	3892054
	4	3903233	3904267
	4	4128333	4129357
	4	4158206	4159230
	4	4176048	4177017
	4	9460948	9461970
	4	9470749	9471709
	4	9485350	9486377
	4	9514518	9515542
	4	9756516	9757429
	4	41725227	41725580
	4	80508265	80509282
	5	101151587	101152484
	5	175433148	175434392
	5	177263414	177264665
	5	180119835	180120763
	5	180166126	180167058
	5	180551360	180552304
	5	180581943	180582887
	5	180794288	180795223
	6	105922	106856
	6	27879027	27880097
	6	27905181	27906178
	6	27925019	27925957
	6	27944928	27945850
	6	28001692	28002639
	6	28014212	28015146
	6	28021009	28021958

	6	28041094	28042014
	6	28423415	28423888
	6	28994433	28995384
	6	29011993	29012952
	6	29039609	29040343
	6	29054087	29055025
	6	29068726	29069655
	6	29079668	29080600
	6	29105660	29106593
	6	29141413	29142348
	6	29149287	29150219
	6	29183013	29183953
	6	29197007	29197997
	6	29230436	29231856
	6	29236254	29237198
	6	29258473	29259427
	6	29274467	29275429
	6	29323010	29323972
	6	29342117	29343064
	6	29364477	29365397
	6	29385057	29385985
	6	29394474	29395418
	6	29407793	29408728
	6	29429547	29430494
	6	29520966	29521937
	6	29541674	29542610
	6	29555722	29556657
	6	132021612	132022541
	6	170948694	170949616
	7	5156720	5157718
	7	6874454	6875425
	7	6906287	6907253
	7	6919253	6920211
	7	97576295	97577301
	7	97595404	97596389
	7	99473688	99474656
	7	141562660	141563619
	7	141586948	141587889
	7	141611204	141611824
	7	141618676	141619617

	7	142723290	142724219
	7	142744167	142744872
	7	142749438	142750376
	7	142759590	142760531
	7	143185558	143186504
	7	143208050	143209004
	7	143632326	143633276
	7	143657064	143658014
	7	143677998	143678925
	7	143701090	143702022
	7	143747495	143748427
	7	143771313	143772242
	7	143774486	143774947
	7	143792201	143793130
	7	143806676	143807629
	7	143815557	143816453
	7	143826206	143827135
	7	143839102	143840031
	7	143854319	143855249
	7	143873833	143874609
	7	143929007	143929936
	7	143947767	143948696
	7	143955792	143956721
	7	143996466	143997395
	7	144015218	144016147
	8	116089	117024
	8	7104141	7105109
	8	7449509	7450477
	8	7562686	7563654
	8	7897501	7898469
	8	11777405	11778304
	8	11786077	11787072
	8	11891126	11892151
	8	12541649	12542674
	8	12553803	12554773
	8	12560561	12561547
	8	21654598	21655408
	9	35859022	35859977
	9	35869463	35870398
	9	35957139	35958095

	9	35991336	35992282
	9	36002909	36003864
	9	36013181	36014127
	9	36021708	36022112
	9	92978296	92979240
	9	92994380	92995405
	9	93504784	93505730
	9	93513500	93514522
	9	107266544	107267500
	9	107288537	107289490
	9	107298054	107299004
	9	107331449	107332408
	9	107352298	107353231
	9	107360741	107361694
	9	107366955	107367908
	9	107379532	107380485
	9	107392035	107393299
	9	107419038	107419997
	9	107456799	107457740
	9	107484616	107485549
	9	114089766	114090713
	9	125239288	125240205
	9	125273081	125274019
	9	125281420	125282358
	9	125288640	125289572
	9	125315491	125316438
	9	125329830	125330756
	9	125370143	125371076
	9	125377017	125377958
	9	125390861	125391814
	9	125423995	125424924
	9	125437409	125438380
	9	125486269	125487201
	9	125512127	125513059
	9	125551212	125552171
	9	125562402	125563349
	10	15028875	15029470
	10	15041102	15042084
	10	15049871	15050878
	10	45753756	45754568

	10	45798887	45799813
	10	135243898	135244822
	10	135294725	135295274
	10	135388657	135389569
	11	86652	87586
	11	3412040	3412979
	11	3620874	3621899
	11	4167546	4168525
	11	4388584	4389525
	11	4399512	4400454
	11	4439290	4440238
	11	4452563	4453459
	11	4470570	4471511
	11	4496039	4496985
	11	4510131	4511072
	11	4536175	4537154
	11	4566421	4567371
	11	4608148	4609092
	11	4615296	4616240
	11	4661021	4661992
	11	4673757	4674710
	11	4682062	4682960
	11	4702982	4703941
	11	4711991	4712932
	11	4730799	4731740
	11	4739842	4740789
	11	4757408	4758308
	11	4773276	4774174
	11	4790212	4791147
	11	4807984	4808932
	11	4814999	4815492
	11	4824666	4825610
	11	4842652	4843641
	11	4853363	4854299
	11	4869470	4870393
	11	4880850	4881794
	11	4897787	4898733
	11	4903130	4904110
	11	4910808	4911826
	11	4928600	4929535

	11	4935952	4936893
	11	4944607	4945569
	11	4958586	4959524
	11	4967392	4968330
	11	4976005	4976943
	11	4994176	4995114
	11	5020213	5021157
	11	5036368	5037306
	11	5058243	5059178
	11	5067756	5068688
	11	5079883	5080857
	11	5090802	5091724
	11	5097324	5098296
	11	5113906	5114838
	11	5125383	5126311
	11	5141897	5142808
	11	5152925	5153872
	11	5172664	5173599
	11	5191584	5192453
	11	5198950	5199840
	11	5220968	5221912
	11	5312697	5313581
	11	5322247	5323176
	11	5335975	5336923
	11	5344592	5345527
	11	5351818	5352575
	11	5363819	5364754
	11	5372738	5373673
	11	5410662	5411606
	11	5423827	5424774
	11	5443431	5444381
	11	5451886	5452832
	11	5461803	5462744
	11	5474719	5475654
	11	5489734	5490589
	11	5509937	5510890
	11	5548453	5549356
	11	5565794	5566735
	11	5572895	5573826
	11	5582197	5583141

	11	5587881	5588846
	11	5602170	5603111
	11	5740520	5741496
	11	5747732	5748691
	11	5757756	5758718
	11	5775971	5776933
	11	5786392	5787276
	11	5798893	5799864
	11	5809087	5810046
	11	5821594	5822551
	11	5841566	5842528
	11	5862189	5863127
	11	5877982	5878920
	11	5895190	5896118
	11	5905523	5906458
	11	5922007	5922987
	11	5924865	5925708
	11	5968577	5969521
	11	5988786	5989724
	11	6007174	6008115
	11	6023284	6024222
	11	6047981	6048922
	11	6067040	6067977
	11	6078519	6079460
	11	6088076	6089022
	11	6129009	6129965
	11	6149840	6150783
	11	6172825	6173779
	11	6190588	6191556
	11	6220454	6221413
	11	6789241	6790188
	11	6806269	6807216
	11	6815959	6816939
	11	6866914	6867864
	11	6890986	6891894
	11	6897852	6898823
	11	6912808	6913731
	11	6942281	6943222
	11	7749947	7750884
	11	7767537	7768456

	11	7794440	7795361
	11	7817524	7818489
	11	7846587	7847519
	11	7870247	7871179
	11	7949268	7950209
	11	7960126	7961067
	11	15009254	15010187
	11	17073548	17074531
	11	29008182	29009045
	11	48238362	48239288
	11	48248981	48249883
	11	48266656	48267564
	11	48285413	48286327
	11	48327775	48328701
	11	48346574	48347479
	11	48366903	48374008
	11	48387040	48388017
	11	48441796	48442723
	11	48453772	48454725
	11	48485617	48486572
	11	48507738	48508583
	11	48510345	48511271
	11	48513273	48514046
	11	48517870	48518810
	11	48533872	48534815
	11	48547616	48548526
	11	48600988	48601903
	11	48611291	48612242
	11	48631615	48632548
	11	48649032	48650012
	11	49919816	49920738
	11	49936664	49937095
	11	49939042	49939929
	11	49941963	49942865
	11	49944602	49945439
	11	49973975	49974901
	11	50003111	50004037
	11	51393463	51394432
	11	51411451	51412395
	11	51425920	51426829

	11	51435459	51436403
	11	51451069	51451981
	11	51455829	51456699
	11	51458687	51459613
	11	51461326	51462182
	11	51483284	51484251
	11	51515282	51516208
	11	51526874	51527800
	11	55086036	55086953
	11	55093232	55094176
	11	55110677	55111660
	11	55135450	55136391
	11	55156005	55156948
	11	55199433	55199843
	11	55211911	55212814
	11	55234247	55235188
	11	55243419	55243904
	11	55245027	55245985
	11	55258632	55259621
	11	55277205	55278138
	11	55304478	55305400
	11	55321945	55322892
	11	55339604	55340533
	11	55370920	55371849
	11	55405813	55406769
	11	55418380	55419312
	11	55432643	55433569
	11	55441068	55442031
	11	55450819	55451745
	11	55482257	55483194
	11	55493675	55494631
	11	55522453	55523430
	11	55540914	55541855
	11	55554441	55555382
	11	55563032	55563973
	11	55578943	55579875
	11	55587106	55588044
	11	55594695	55595627
	11	55606228	55607211
	11	55623068	55623997

	11	55670820	55671622
	11	55681129	55682058
	11	55702935	55703876
	11	55715595	55716399
	11	55724691	55725105
	11	55735037	55735939
	11	55746590	55747582
	11	55761160	55762101
	11	55782493	55783433
	11	55797895	55798866
	11	55821841	55822776
	11	55838587	55839522
	11	55850277	55851212
	11	55860784	55861713
	11	55864720	55864860
	11	55872519	55873454
	11	55884115	55884732
	11	55889849	55890784
	11	55900142	55900970
	11	55904250	55905194
	11	55909991	55910862
	11	55926873	55927793
	11	55932976	55933397
	11	55944094	55945029
	11	55955807	55956763
	11	55978332	55979296
	11	55999585	56000538
	11	56019676	56020695
	11	56043181	56044092
	11	56057606	56058538
	11	56064016	56064946
	11	56085783	56086718
	11	56102618	56103544
	11	56113515	56114471
	11	56127723	56128670
	11	56143085	56144026
	11	56149111	56150065
	11	56161205	56162124
	11	56180172	56181156
	11	56184737	56185708

	11	56216144	56217068
	11	56229948	56230877
	11	56237053	56237973
	11	56246967	56247902
	11	56257914	56258846
	11	56267775	56268746
	11	56279710	56280646
	11	56294043	56294978
	11	56309819	56310733
	11	56344253	56345197
	11	56364902	56365838
	11	56380034	56380978
	11	56387222	56388286
	11	56396408	56397375
	11	56400639	56401578
	11	56408968	56409918
	11	56431162	56432091
	11	56436496	56437420
	11	56467864	56468778
	11	56507655	56508568
	11	56510307	56511242
	11	56518473	56519460
	11	56542741	56543682
	11	56557597	56558577
	11	56569335	56570270
	11	56587052	56587992
	11	56738525	56739451
	11	56756389	56757315
	11	56785566	56786493
	11	56796862	56797197
	11	56805009	56805933
	11	56812554	56812884
	11	56823175	56823608
	11	57633774	57634752
	11	57684774	57685702
	11	57713027	57713943
	11	57798425	57799375
	11	57844821	57846077
	11	57876192	57877133
	11	57885975	57886916

	11	57911898	57912893
	11	57938561	57939193
	11	57946917	57947846
	11	57957963	57958904
	11	57970679	57971614
	11	57982256	57983191
	11	57995391	57996347
	11	58034416	58035330
	11	58059301	58060243
	11	58085112	58085945
	11	58111286	58112143
	11	58116461	58117101
	11	58125601	58126542
	11	58133303	58134236
	11	58155044	58155959
	11	58169941	58170882
	11	58189808	58190734
	11	58206683	58207624
	11	58274652	58275578
	11	59077115	59078061
	11	59100229	59101558
	11	59131932	59132864
	11	59158826	59159764
	11	59189455	59190426
	11	59210642	59211586
	11	59224434	59225375
	11	59244903	59245835
	11	59259125	59260083
	11	59271049	59271981
	11	59282386	59283327
	11	59299170	59300183
	11	59480392	59481318
	11	59495992	59496965
	11	59508798	59509522
	11	59516284	59517047
	11	67489930	67490955
	11	67503068	67504018
	11	67741706	67742728
	11	71304440	71305459
	11	71331059	71332078

	11	71604453	71605484
	11	71614231	71615165
	11	72960360	72960913
	11	74782165	74783136
	11	74799799	74800758
	11	74842148	74842917
	11	86543570	86544513
	11	86568076	86569053
	11	105064817	105065663
	11	123624291	123625226
	11	123676119	123677057
	11	123711666	123712611
	11	123732389	123733231
	11	123777139	123778080
	11	123810324	123811277
	11	123813577	123814545
	11	123847406	123848365
	11	123864873	123865868
	11	123880376	123881185
	11	123886282	123887214
	11	123893720	123894652
	11	123900330	123901262
	11	123908776	123909708
	11	123925500	123926433
	11	123964306	123965239
	11	123975759	123976681
	11	124029128	124030060
	11	124055977	124056912
	11	124077890	124078852
	11	124085456	124086391
	11	124095437	124096383
	11	124109122	124110318
	11	124120423	124121334
	11	124134828	124135760
	11	124179739	124180662
	11	124189161	124190093
	11	124194739	124195683
	11	124208658	124209630
	11	124227447	124228395
	11	124235518	124236447

	11	124246773	124247703
	11	124252301	124253239
	11	124266309	124267247
	11	124274129	124274997
	11	124293841	124294767
	11	124310049	124310981
	11	124330460	124331422
	11	124350896	124351824
	11	124386206	124387138
	11	124401502	124401880
	11	124412621	124413550
	11	124440016	124440942
	11	124471498	124472154
	12	8567376	8568371
	12	8580258	8581210
	12	8589999	8590979
	12	46986356	46987088
	12	48596125	48597075
	12	48749442	48750328
	12	48779134	48780088
	12	48788266	48789169
	12	48835816	48836730
	12	48919415	48920350
	12	48953665	48954597
	12	49021623	49022595
	12	52501102	52502047
	12	55509317	55510255
	12	55523619	55524557
	12	55552631	55553584
	12	55587645	55588600
	12	55614809	55615756
	12	55641072	55642007
	12	55656323	55657218
	12	55677534	55678467
	12	55688075	55689016
	12	55705619	55706541
	12	55714384	55715319
	12	55725485	55726417
	12	55736622	55737560
	12	55758895	55759830

	12	55770585	55771520
	12	55782229	55783161
	12	55790060	55791057
	12	55794313	55795248
	12	55820038	55820973
	12	55845998	55846933
	12	55862987	55863922
	12	55886147	55887097
	12	55916377	55917255
	12	55945011	55945937
	12	55968199	55969125
	12	56005407	56006338
	12	56030676	56031614
	12	56039979	56040447
	13	42005400	42006394
	13	42013973	42014998
	13	42016806	42017799
	13	47034670	47035343
	13	64316447	64316993
	13	64411061	64411806
	13	68476376	68477356
	13	68485133	68486080
	14	19377594	19378571
	14	19806453	19807382
	14	20181098	20182042
	14	20201505	20202452
	14	20215587	20216525
	14	20228069	20228944
	14	20248482	20249420
	14	20264495	20265443
	14	20295608	20296528
	14	20315526	20316464
	14	20336359	20337302
	14	20344427	20345368
	14	20373864	20374794
	14	20388766	20389734
	14	20403826	20404758
	14	20424706	20425713
	14	20443750	20444721
	14	20470322	20471252

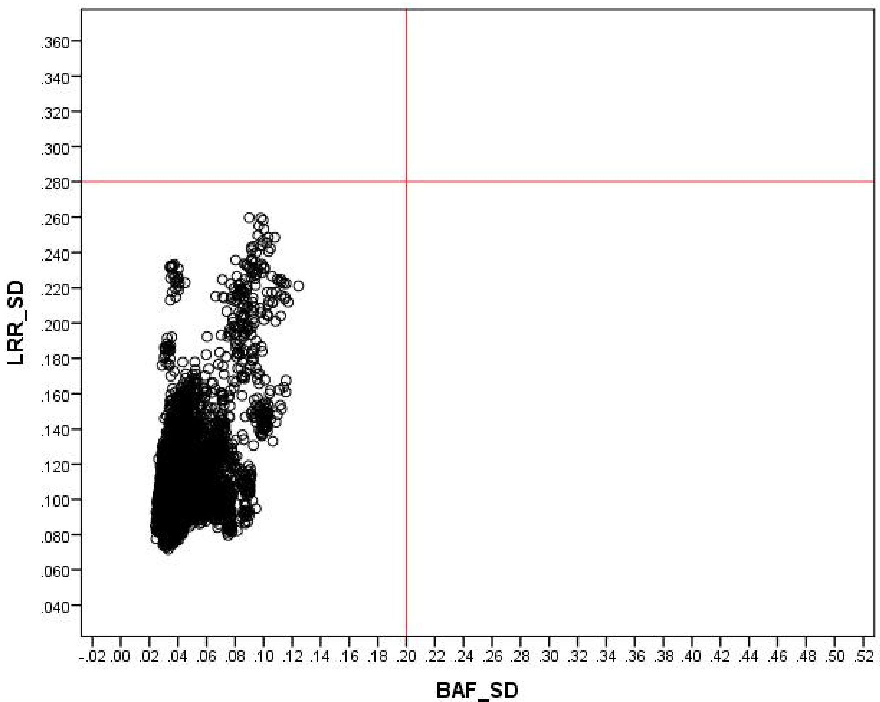
	14	20482423	20483352
	14	20502006	20502917
	14	20512075	20513014
	14	20528204	20529139
	14	20561514	20562434
	14	20585659	20586594
	14	20611895	20612818
	14	20647387	20648337
	14	20652084	20652457
	14	20665597	20666529
	14	20677456	20678354
	14	20691938	20692858
	14	20697561	20698502
	14	20710981	20711922
	14	21108858	21109850
	14	21623099	21624031
	14	22037937	22038875
	14	22070035	22070976
	14	22102069	22102998
	14	22133297	22134235
	14	22138204	22139150
	14	23102676	23103716
	14	23171352	23172277
	14	25826970	25827973
	14	52223708	52224703
	14	52229812	52230758
	14	52238506	52238814
	15	21165990	21166935
	15	21174579	21175538
	15	22177219	22178164
	15	22185810	22186769
	15	22268237	22269196
	15	22297501	22298447
	15	22318854	22319782
	15	22332368	22333348
	15	22344660	22345590
	15	22368576	22369514
	15	22382473	22383420
	15	22413462	22414382
	15	102345923	102346858

	15	102358390	102359325
	15	102368944	102369879
	15	102388848	102389784
	15	102416167	102417101
	15	102462348	102463262
	15	102466975	102467844
	15	102477753	102478696
	16	3254247	3255182
	16	3265566	3266546
	16	3405941	3406876
	17	2965966	2966901
	17	2995355	2996290
	17	3019721	3020697
	17	3029907	3030845
	17	3057187	3058146
	17	3100813	3101739
	17	3118915	3119841
	17	3143970	3144905
	17	3168905	3169856
	17	3181267	3182211
	17	3194932	3195876
	17	3213605	3214648
	17	3289224	3290165
	17	3300763	3301704
	17	3323880	3324824
	17	3336167	3337135
	17	56232515	56233444
	17	56247017	56247937
	18	14570843	14571775
	18	14613209	14614162
	19	94062	95005
	19	104601	105469
	19	110679	111593
	19	156282	157216
	19	8841391	8842332
	19	9193612	9194074
	19	9203921	9204859
	19	9212948	9213919
	19	9225507	9226439
	19	9232760	9233163

	19	9236691	9237626
	19	9296458	9297393
	19	9301912	9302831
	19	9314928	9315908
	19	9324578	9325513
	19	9345820	9346752
	19	9361720	9362736
	19	9371653	9372473
	19	9376069	9377045
	19	9389109	9390045
	19	14909989	14910948
	19	14938097	14939053
	19	14951763	14952689
	19	14962847	14963756
	19	14974601	14975584
	19	14991241	14992167
	19	14997658	14998084
	19	15002671	15003616
	19	15014054	15014985
	19	15027195	15028213
	19	15037864	15038869
	19	15052301	15053257
	19	15197877	15198821
	19	15251919	15252819
	19	15838854	15839798
	19	15852203	15853150
	19	15904859	15905803
	19	15917894	15918847
	19	16059818	16060765
	19	16162787	16163525
	19	57272232	57272600
	21	14916552	14917506
	21	14953393	14954325
	21	33993371	33994353
	22	16448827	16449771

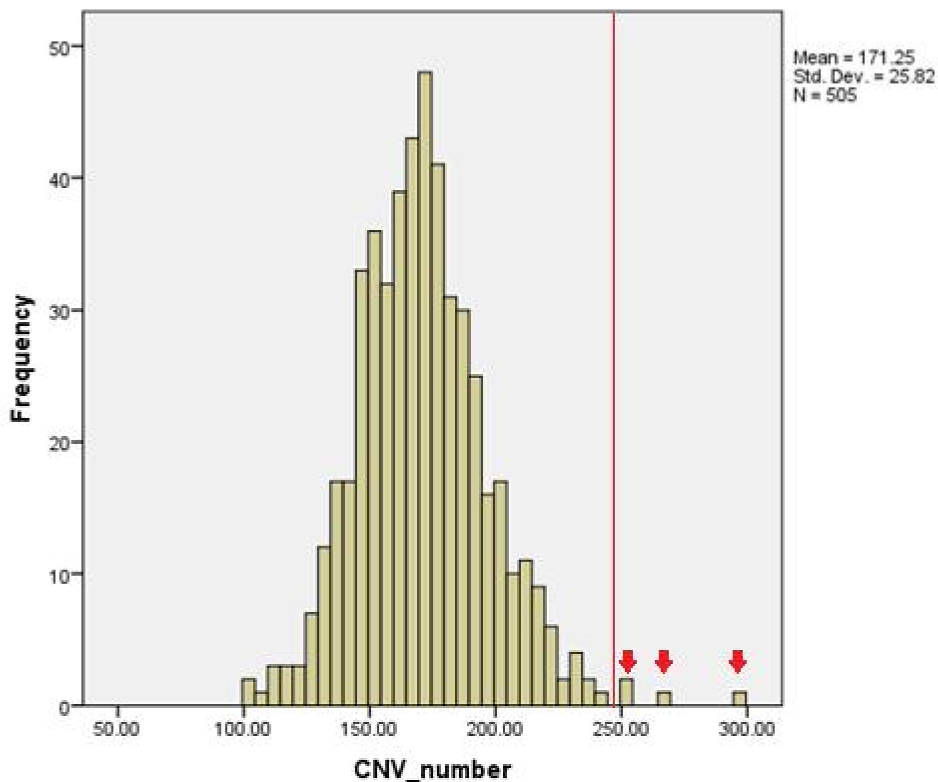
Appendix F: Tables and figures related to the QuantiSNP and PennCNV QC analysis results.

F. 1: QuantiSNP subject QC filtering based on the LRR_SD and BAF_SD thresholds.



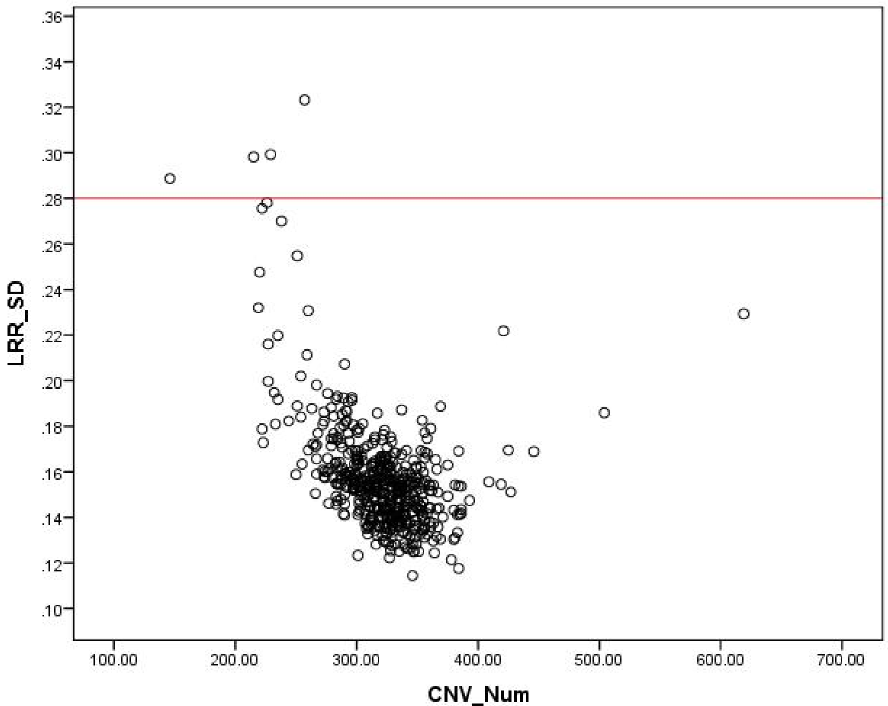
LRR_SD: Log R Ratio standard deviation, **BAF_SD:** B Allele Frequency standard deviation. This figure illustrates that the data of all patient satisfied the LRR_SD and BAF_SD criteria.

F. 2: QuantiSNP subject QC filtering based on the number of predicted variations.



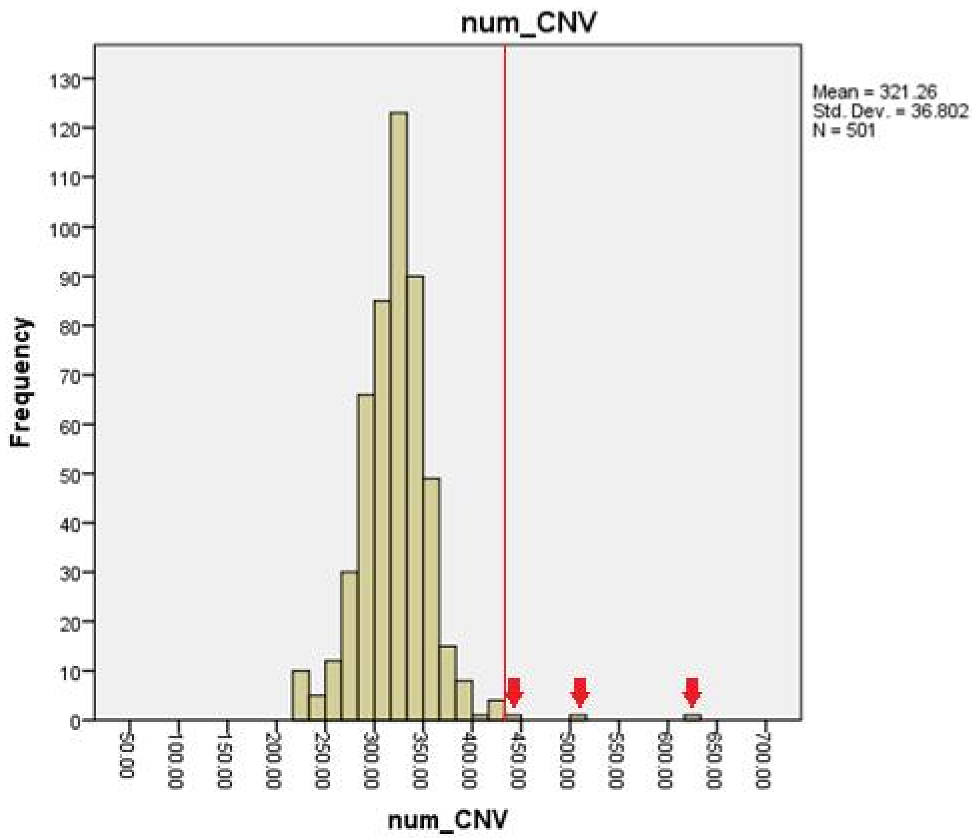
Mean variation number + 3 SD = 171.25 + (3 * 25.82) = 248.71. Three individuals had excessive number of predicted variations (shown by red arrows). As a result, these patients were excluded from the QuantiSNP results.

F. 3: PennCNV subject QC filtering based on the LRR_SD criterion.



LRR_SD: Log R Ratio standard deviation, **CNV_Num:** The number of INDELs/CNVs. Four individuals had LRR_SD > 0.28, and were excluded from the PennCNV data.

F. 4: PennCNV QC filtering based on the number of predicted variations.



Mean variation number + 3 SD = 321.26 + (3 * 36.80) = 431.66. Three patients with excessive number of predicted variations (indicated by red arrows) were removed from the PennCNV results.

F. 5: The number of patients and the features of INDELs/CNVs that passed the QC filtering of QuantiSNP and PennCNV algorithms.

Variable	**QuantiSNP**	**PennCNV**
Number of patients	501	497
Total predicted INDELs/CNVs in the cohort	85,469	159,050
Average number of INDELs/CNVs per individual	170.60	320.02

Type	**N**	**%**	**N**	**%**
INDELs	17,026	19.9	32,564	20.5
CNVs	68,443	80.1	126,486	79.5

INDELs/CNVs per CN State		**N**	**%**	**N**	**%**
(CN= 0)	Two copy deletion	52,898	61.9	51,264	32.2
(CN= 1)	One copy deletion	24,317	28.5	79,661	50.1
(CN= 3)	One copy duplication	4,537	5.3	27,555	17.3
*(CN= 4,5)	Two or more copy duplication	3,717	4.35	570	0.4

N: Number, **CN**: Copy number state. *Please note that QuantiSNP assigns the CN state 4 for variants that exist in 4 copies and CN state 5 for variants that exist in 5 or more copies in a genome. However, PennCNV assigns the CN state 4 for variants that exist in 4 or more copies in a genome.

F. 6: The baseline features of the 495 colorectal cancer patients.

Features	Number	%
Sex		
Female	194	39.19
Male	301	60.81
Age at diagnosis	median: 61.4 years (range: 20.7-75 years)	
<65	301	60.81
≥65	194	39.19
Location		
Colon	328	66.26
Rectum	167	33.74
Histology		
Non-mucinous	438	88.48
Mucinous	57	11.52
Stage		
I	89	17.98
II	193	38.99
III	164	33.13
IV	49	9.90
Grade		
Well/moderately differentiated	457	92.32
Poorly differentiated	34	6.87
Unknown	4	0.81
Vascular invasion		
Absent	300	60.61
Present	158	31.92
Unknown	37	7.47
Lymphatic invasion		
Absent	290	58.59
Present	166	33.54
Unknown	39	7.88
Familial risk		
Low risk	244	49.29
Moderate/high risk	251	50.71
MSI status		
MSI-L/MSS	421	85.05
MSI-H	53	10.71
Unknown	21	4.24

Tumour *BRAF* Val600Glu mutation		
Absent	402	81.21
Present	47	9.49
Unknown	46	9.29
Colorectal cancer cases		
Sporadic cases	465	93.94
Lynch syndrome cases	14	2.83
FCCX cases	13	2.62
FAP cases	3	0.61

MSI-H: microsatellite instability-high; MSI-L: microsatellite instability-low, and MSS: microsatellite stable; FCCX: familial colorectal cancer type X.

Appendix G: Summary statistics of INDELs/CNVs that were predicted by both QuantiSNP and PennCNV algorithms.

Variable	**Number**
Number of patients	495
Males	301
Females	194
Total INDELs/CNVs in the cohort	74,261
Average number of INDELs/CNVs per individual	150.02

Type	**N**	**%**
INDELs	14,642	19.72
CNVs	59,619	80.28

INDELs/CNVs per CN State		**N**	**%**
(CN= 0)	Two copy deletion	48,071	64.73
(CN= 1)	One copy deletion	23,168	31.20
(CN= 3)	One copy duplication	2,810	3.784
(CN= 4)	Two or more copy duplication	212	0.28

N: Number, **CN**: Copy number state.

Appendix H: Number of CNVs per patient categories.

Variable category		number of patients	Mean	SD	P-Value
Sex	Female	194	139.41	22.49	0.49
	Male	301	140.83	22.09	
Age at diagnosis	<65	312	138.67	21.83	0.036
	>=65	183	143.01	22.71	
Histology	Non-mucinous	438	140.89	22.45	0.09
	mucinous	57	135.58	20.1	
Location	colon	328	140.12	21.79	0.83
	rectum	167	140.58	23.16	
Stage	I	89	139.62	24.12	(I,II) 0.72; (I,III) 0.81; (I,IV) 0.98; (II,III) 0.89; (II,IV) 0.78; (III,IV) 0.86
	II	193	140.67	22.07	
	III	164	140.34	22.56	
	IV	49	139.71	18.54	
Grade	well/moderately differentiated	461	140.25	22.51	0.92
	poorly differentiated	34	140.65	18.33	
Vascular _invasion	no invasion	337	141.13	22.97	0.21
	invasion	158	138.46	20.54	
Lymphatic _ invasion	no invasion	329	141.22	23.19	0.19
	invasion	166	138.42	20.16	
MSI status	MSI-L/MSS	442	140.29	22.77	0.96
	MSI-H	53	140.13	17.38	
Familial risk	Low risk	244	141.20	22.77	0.36
	Moderate/high risk	251	139.38	21.72	